Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

Facultad de Ingeniería de Sistemas e Informática

Escuela Profesional de Ingeniería de Software

Sistema inteligente basado en Machine Learning para la detección de fraude de facturación de agua potable

TESIS

Para optar el Título Profesional de Ingeniero de Software

AUTOR

Anthony Joffre CARRILLO ROSALES

ASESOR

David Santos MAURICIO SÁNCHEZ

Lima, Perú

2019

Referencia bibliográfica

Carrillo, A. (2019). *Sistema inteligente basado en Machine Learning para la detección de fraude de facturación de agua potable*. Tesis para optar el título profesional de Ingeniero de Software. Escuela Profesional de Ingeniería de Software, Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional Mayor de San Marcos, Lima, Perú.

HOJA DE METADATOS COMPLEMENTARIOS

Código Orcid del autor:

Código Orcid del asesor o asesores:

0000-0001-9262-626X

DNI del autor:

46398320

Grupo de Investigación:

Facultad de Ingeniería de Sistemas e Informática

Institución que financia parcial o totalmente la investigación:

Universidad Nacional Mayor de San Marcos

Ubicación geográfica donde se desarrolló la investigación. Debe incluir localidades y coordenadas geográficas.

Cercado de Lima, LIMA, PERÚ

Coordenadas: -12.053427, -77.085709

Año o rango de años que la investigación abarco:

2018 - 2019

Universidad Nacional Mayor de San Marcos

Universidad del Perú, DECANA DE AMÉRICA

Facultad de Ingeniería de Sistemas e Informática

Escuela Profesional de Ingeniería de Software

Acta de Sustentación de Tesis

Siendo las 20:15 del día 23 de octubre del año 2019, se reunieron los docentes designados como miembros de Jurado de la Tesis, presidido por el Dr. Javier Arturo Gamboa Cruzado, Mg. Juan Gamarra Moreno (Miembro), y el Dr. David Santos Mauricio Sánchez (Miembro Asesor) para la sustentación de Tesis intitulada: **"SISTEMA INTELIGENTE BASADO EN MACHINE LEARNING PARA LA DETECCIÓN DE FRAUDE DE FACTURACIÓN DE AGUA POTABLE"**; por el bachiller **Anthony Joffre Carrillo Rosales**, para optar el Título Profesional de Ingeniero de Software.

Acto seguido de la exposición de la Tesis, el Presidente invitó al bachiller a dar respuesta a las preguntas establecidas por los Miembros de Jurado.

El bachiller en el curso de sus intervenciones demostró pleno dominio del tema, al responder con acierto y fluidez a las observaciones y preguntas formuladas por los señores miembros del Jurado.

Finalmente habiéndose efectuado la calificación correspondiente por los miembros de Jurado, el bachiller obtuvo la nota de 18. (En letras) dieciocho

A continuación el Presidente del Jurado, Dr. Javier Arturo Gamboa Cruzado **Ingeniero de Software**.

Siendo las 21:00 horas, se levantó la sesión.

.....
Dr. Javier Arturo Gamboa Cruzado
Presidente

.....
Mg. Juan Gamarra Moreno
Miembro

.....
Dr. David Santos Mauricio Sánchez
Miembro Asesor

Quiero dedicar este trabajo a mis padres, Santa Susana y Job Vigil.
A mi madre quien me refleja siempre su infinita fortaleza, a mi padre por su apoyo incondicional, y a ambos por ser mi motivo para seguir creciendo en mi carrera profesional.

AGRADECIMIENTOS

Al profesor Dr. David Mauricio Sánchez, por su apoyo, orientación, esfuerzo y consejos desinteresados, gracias a los cuales fue posible culminar este trabajo de investigación y cumplir con los objetivos trazados.

Al Dr. Javier Gamboa Cruzado y al Mg. Juan Gamarra Moreno, les agradezco por sus comentarios, sugerencias al presente trabajo de tesis y ser miembros del jurado evaluador.

A Richard Palomino y Mijail Rivera, por haberme facilitado los datos de consumo de agua de los clientes de la Municipalidad de Gaza, en Palestina.

A la Universidad Nacional Mayor de San Marcos y a la Facultad de Ingeniería de Sistemas e Informática, por haberme permitido realizar y concluir mis estudios superiores.

A cada uno de mis profesores en el pregrado, quienes ciclo a ciclo me inculcaron no solo conocimientos acádemicos, sino la perseverancia, el poder superarme como persona, gracias por haberme guiado en mi desarrollo profesional.

A mi familia, en especial a mis padres por su gran apoyo incondicional, por transimitirme su fortaleza y perseverancia que admiro de ellos.

A mi hermana Helen, por echarme una mano cuando siempre lo necesité, por aportar considerablemente en mi proyecto.

A mi hermana Magaly Camila, que en paz descanse, quien desde el cielo me está cuidando. Gracias a ellos porque sé que siempre desearán lo mejor para mí.

A mis amigos que conocí en la etapa de pregrado, quienes me demostraron que, con mucha paciencia, perseverancia y trazándose un objetivo, es posible de hacer realidad todo lo que uno se proponga, gracias a ellos porque indirectamente me ayudaron a culminar este trabajo.

Sistema Inteligente Basado en Machine Learning para la Detección de Fraude de Facturación de Agua Potable

RESUMEN

En la actualidad no existe una herramienta o un sistema el cual compruebe con gran exactitud (al menos de un 97 %) la detección de usuarios que cometen fraude en la facturación del consumo de agua potable, ya sea por conexiones ilícitas o adulteración de sus medidores de agua. Sin embargo, en el trabajo de investigación titulado Sistema Inteligente para detectar fraude en el servicio de Agua Potable de una Empresa Sanitaria (Palomino y Rivera, 2016) se obtuvo una tasa de 95.7 % de exactitud en la detección de fraude en Gasa, Palestina. Cabe resaltar que la cantidad de pérdida económica es sumamente considerable, así que la creación de una herramienta o sistema para detectar a estos usuarios fraudulentos es de bastante importancia para las empresas generadoras de agua potable. En el presente trabajo de investigación se propone desarrollar un Sistema Inteligente basado en un modelo híbrido de técnicas de minería de datos que pretende mejorar la tasa de exactitud en detección de un cliente en fraude de facturación de agua potable. Para el entrenamiento y la validación del modelo híbrido se pretende usar un dataset histórico del consumo de agua de los clientes de una empresa sanitaria en Palestina, así se obtendrá una tasa de 97.71 % de exactitud de detección de fraude.

Palabras clave: Fraude de facturación de agua, Minería de Datos, Clasificador por votación, Máquina de Soporte Vectorial, Regresión linear, Árbol de decisión.

Intelligent System based on Automatic Learning for the Detection of Drinking Water Billing Fraud

ABSTRACT

Currently, there is no tool or system that verifies with great accuracy (at least 97%) the detection of users who commit fraud in the billing of drinking water consumption, either through illicit connections or adulteration of their meters of water. However, in the research work entitled Intelligent System to detect fraud in the Drinking Water service of a Healthcare Company (Palomino and Rivera, 2016), a 95.7% accuracy rate was obtained in fraud detection in Gaza, Palestine. It should be noted that the amount of economic loss is extremely considerable, so the creation of a tool or system to detect these fraudulent users is quite important for companies generating drinking water. In this research work we propose to develop an Intelligent System based on a hybrid model of data mining techniques that aims to improve the accuracy rate in detecting a customer in drinking water billing fraud. For the training and validation of the hybrid model, we intend to use a historical dataset of the water consumption of the clients of a health company in Palestine, thus obtaining a rate of 97.71% accuracy of detection of fraud.

Keywords: Water billing fraud, Data Mining, Voting Sorter, Vector Support Machine, Linear Regression, Decision Tree.

TABLA DE CONTENIDOS

Lista de Figuras

Lista de Tablas

CAPÍTULO 1: INTRODUCCIÓN

1.1. Antecedentes

El agua es un recurso muy necesario en la humanidad. Según un informe de la ONU (ONU, 2010), 828 millones de personas habitan asentamientos informales, los cuales están esparcidos alrededor del mundo y es un gran desafío el de abastecer a estas personas de agua potable y servicios sanitarios. El acceso al agua potable se mide por el número de personas que pueden obtener agua potable con razonable facilidad, expresado como porcentaje de la población total. Es un indicador de la salud de la población en un país y capacidad del mismo de conseguir agua, purificarla y distribuirla. El agua es esencial para la vida, sin embargo, más de mil millones de personas carecen de acceso a agua potable y casi 2 millones de personas carecen de acceso a servicios de saneamiento. La mayoría de esas personas viven en países de ingreso bajo y mediano (CivilGeek, 2011). Además, según (ONU, 2010), los motivos más importantes son:

- Entre 250 y 500 millones de metros cúbicos de agua potable es desperdiciada cada año en varias ciudades del mundo.

- No tener que desperdiciar esta cantidad de agua ayudaría a proveer de agua potable a unos 15 a 20 millones de personas adicionales en cada gran ciudad.

Uno de los grandes problemas que encuentran las empresas de abastecimiento de agua dentro del consumo de agua potable es el fraude en la facturación del agua, lo cual trae como consecuencia el hecho de no lograr cobrar en un cliente final el consumo total de su facturación, este índice es alrededor de un 30 % (Patiño, 2014) del agua no facturada por intervenciones de las personas en sus propios medidores, con el objetivo de

disminuir la lectura de m^3. Teniendo un estimado de 7000.000 m^3, lo que en dinero es un aproximando de $468 MM.

Para atacar este problema existen diversos tipos de técnicas con determinada exactitud, como el árbol de decisión con una exactitud del 75 % (Patiño, 2014), Coeficiente de Paerson (Monedero, 2015), algoritmos estadísticos (Biscarri, 2015), detección de datos atípicos (De Castro, 2015), entre otras. Sin embargo, existe un problema localizado en las distintas empresas de abastecimiento de agua (Patiño, 2014).

1.2. Problema

El problema de la detección de fraude en la facturación de agua es determinar los usuarios que manipulan sus medidores de agua o realizan instalaciones paralelas con el fin de pagar menos en su facturación; como consecuencia disminuir las pérdidas económicas de las empresas de abastecimiento de agua por el consumo de agua no facturada.

1.3. Importancia del problema

El fraude en la facturación de agua, ocurre en diversos países en el mundo, generando grandes pérdidas económicas. A continuación, veremos los casos de Chile y Perú.

La empresa chilena Aguas Andinas, obtuvo una producción promedio mensual en el año 2014 de 58 MM m^3, de los cuales se cobró 40 MM m^3, es decir, se dejó de facturar un promedio de 18 MM de m^3 al mes.

Según estudios realizados por SUEZ, la repercusión que tiene cada factor en el agua no contabilizada es:

Factor	Porcentaje (%)	Magnitud de MM m^3 al mes
Pérdida técnica	57	10.33
Pérdida micro medición	33	5.9
Pérdida uso irregular	10	1.8

Tabla 1.1. Desglose del agua mensual no facturada (Ledesma O., 2014)

La pérdida por el uso irregular incluye robo de grifos, servicios no enrolados, conexiones paralelas e intervención en sus medidores. Aguas Andinas estimó que de los 1,8 MM m^3 de agua potable perdidos al mes por uso irregular, cerca de 700,000 m^3 corresponden a clientes enrolados con instalaciones irregulares (conexión paralela o intervención en sus medidores). Si se considera que, al no facturar dicho volumen, se deja de cobrar por concepto de agua potable, así se obtendrá una pérdida económica mensual que bordea los $468 MM (Patiño, 2014).

El presente trabajo se encargará de evaluar el tercer y cuarto punto del factor “perdida por uso irregular”, vale decir, la detección de clientes que facturan un volumen menor al correspondiente a su consumo real, ya sea por conexiones paralelas o por intervención en sus medidores.

Según parte de la metodología propuesta por la IWA (International Water Association) propuso una forma estandar para contabilizar el ANF (Agua no Facturada), representada en la Figura 1.1. donde se realiza un balance ordenado del volumen de entrada al sistema de conducciones y de distribución de agua potable.

<table>
<tr><td rowspan="6">Agua potable producida en fuentes propias
(m^3/año)</td><td rowspan="11">Volumen de entrada al Sistema de Distribución
(m^3/año)</td><td rowspan="2">Agua exportada</td><td rowspan="6">Consumos autorizados
(m^3/año)</td><td rowspan="3">Consumos autorizados facturados</td><td>Agua exportada facturada</td><td rowspan="3">Agua facturada (m^3/año)</td></tr>
<tr><td rowspan="2">Consumos facturados medidos</td></tr>
<tr><td rowspan="2">Agua suministrada
(m^3/año)</td></tr>
<tr><td>Consumos facturados no medidos.</td></tr>
<tr><td rowspan="2">Consumos autorizados no facturados</td><td>Consumos no facturados medidos</td><td rowspan="7">Agua no facturada ANF (*) (m^3/año)</td></tr>
<tr><td>Consumos no facturados no medidos</td></tr>
<tr><td rowspan="5">Agua potable importada.
(m^3/año)</td><td rowspan="5">Pérdidas
(m^3/año)</td><td rowspan="2">Pérdidas comerciales</td><td>Consumos no autorizados</td></tr>
<tr><td>Imprecisión en la medición</td></tr>
<tr><td rowspan="3">Pérdidas físicas</td><td>Pérdidas en conducciones y redes</td></tr>
<tr><td>Pérdidas y reboses de estanques</td></tr>
<tr><td>Pérdidas de arranques (Matriz -MAP)</td></tr>
<tr><td colspan="7">Notas: * Agua no facturada (ANF): En inglés IWA define esta parte del balance como "Non Revenue Water" (NRW), lo que se traduciría "Agua sin beneficio o agua no rentabilizada". IWA recomienda no usar el término Unaccoundted for Water (UFW), que se parece al español "Agua No contabilizada (ANC)". Si se debería calcularse según el desglose inclinado en la tabla</td></tr>
</table>

Figura 1.1. Esquema del Balance de Agua de la IWA (ECONSSA Chile, 2014)

Un antecedente relevante para el presente estudio lo constituye el estudio denominado "Estudio Caracterización de Aguas no Contabilizadas en Empresas de Servicios Sanitarios", realizado por INECON para la Superintendencia de Servicios Sanitarios (SISS) durante el año 2012 a la emprea chilena ECCONSA y se muestra en la siguiente Figura 1.2.

<table>
<tr><td rowspan="8">Volumen de entrada al Sistema de Distribución (m^3/año)

40,000.000

(m^3/año)</td><td rowspan="4">Consumos autorizados

35,840.000
(m^3/año)</td><td rowspan="2">Consumos Autorizados facturados

35,500.000
(m^3/año)</td><td>Consumos facturados medidos

35,000.000</td><td rowspan="2">Agua facturado

35,500.000
(m^3/año)</td></tr>
<tr><td>Consumos facturados no medidos.

500.000</td></tr>
<tr><td rowspan="2">Consumos autorizados No facturados

340.000
(m^3/año)</td><td>Consumos no facturados no medidos

120.000</td><td rowspan="6">Agua no facturada ANF

4,500.000
(m^3/año)</td></tr>
<tr><td>Consumos no facturados no medidos

220.000</td></tr>
<tr><td rowspan="4">Pérdidas

4,160.000
(m^3/año)</td><td rowspan="2">Pérdidas comerciales

2,150.000
(m^3/año)</td><td>Consumos no autorizados

50.000</td></tr>
<tr><td>Imprecisión en la medición

2,100.00</td></tr>
<tr><td rowspan="3">Pérdidas físicas

2,010.000
(m^3/año)</td><td>Pérdidas en conducciones y redes

1,005.000</td></tr>
<tr><td>Pérdidas y reboses de estanques

0</td></tr>
<tr><td>Pérdidas en arranques (matriz-MAP)

1,005.000</td></tr>
</table>

Figura 1.2. Planilla para el cálculo de balance de agua (ECONSSA Chile, 2014)

Un caso en Péru la empresa Sedapal, se aprecia que en los últimos 6 años (2013-2018), se ha producido cierta cantidad de volumen anualmente, alcanzando una producción promedio anual de 699 millones de m^3 y asu vez teniendo un porcentaje de agua no facturada (ANF) por cada año, ocasionando una pérdida de millones de soles (Sedapal, 2018). A continuación, se muestra en la Figura 1.3. el porcentaje de la cantidad de agua no facturada desde el año 2013 hasta el 2018.

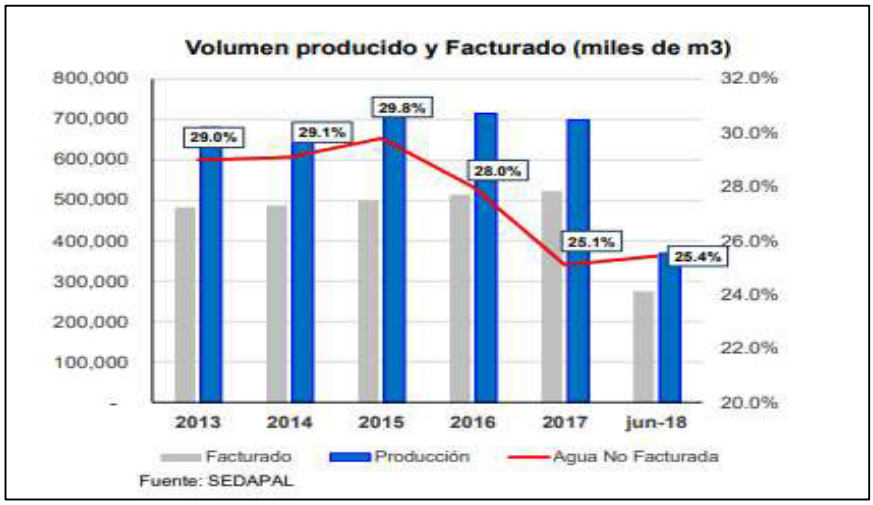


Figura 1.3. Producción y agua no facturada (Sedapal, 2018)

1.4. Motivación

En la actualidad existen muchas empresas prestadoras de servicios de saneamiento como agua potable y alcantarillado sanitario, los cuales tienen como objetivo abastecer a toda la población. Actualmente existe un porcentaje de personas que no cuenta con este servicio por diversos motivos, debido a ello se necesita hacer el uso adecuado del recurso que brindan, en virtud de la creciente demanda. Por dicho motivo se requiere un control de calidad de las personas que cuentan con el servicio, ya que pueden incurrir en algún tipo de anomalías (posible motivo: robo), lo cual ocasiona una pérdida económica a las empresas prestadoras del servicio de saneamiento de agua potable, con una cantidad de agua no facturada, donde los ingresos mensuales son menores a los reales, esto quiere decir agua no facturada.

Para esto se requerirá de un Software que pueda aprender automáticamente a identificar a los clientes que cometen este tipo de fraude en un tiempo corto de consumo, lo cual sería más eficiente que una inspección manual en cada una de las viviendas, pues esto demandaría una gran inversión para la empresa prestadora del servicio de saneamiento de agua, además, realizarla de esta forma demandaría mucho tiempo en detectar donde se cometiese el fraude, para ello ya no existiría una inversión para tomar esta medida en los sectores donde no llegan este servicio por parte de la empresa de Servicio de Agua.

1.5. Objetivos

1.5.1. Objetivo general

Desarrollar un Sistema Inteligente basado en Árboles de decisión, Regresión Logística Lineal y Máquina de Soporte Vectorial para mejorar la exactitud en la detección de fraude

por el consumo de agua no facturada mensualmente, que son cometidos por usuarios de una Empresa Prestadora de servicio de Saneamiento de Agua Potable.

1.5.2. Objetivos específicos

O.E.1. Revisar la literatura para identificar variables necesarias para la detección de fraude de facturación de agua.

O.E.2. Diseñar un modelo híbrido basado en técnicas de minería de datos.

O.E.3. Implementar un Sistema Inteligente que validará el dataset de la literatura mencionada anteriormente.

O.E.4. Realizar pruebas numéricas con un dataset de la literatura para determinar la exactitud de los resultados del sistema propuesto.

1.6. Propuesta

Se propone el desarrollo de un sistema basado en machine learning, el cual será capaz de detectar a los usuarios que están cometiendo este delito para su facturación de agua potable. La idea planteada para la propuesta se presenta en la siguiente figura.

Figura 1.4. Idea básica propuesta

1.7. Organización de la tesis

El presente trabajo de tesis está compuesto por seis capítulos, que a continuación se describe sucintamente.

En el capítulo 2, se describe el estado del arte, donde se dan a conocer investigaciones respecto a la detección de fraude en la facturación de agua y sus porcentajes de exactitud en la detección.

Luego, en el capítulo 3 se presentará el aporte, donde se identificará la metodología, el desarrollo y la realización de la propuesta para el proyecto de tesis, así como la descripción de la herramienta.

En el capítulo 4 se describirá el desarrollo del Sistema Inteligente en mención, la arquitectura, el modelado del sistema, la validación del sistema y los requerimientos para el sistema.

A continuación, en el capítulo 5 se presentará el entrenamiento y la validación del modelo inteligente sobre un dataset, así se mostrará los respectivos resultados del entrenamiento.

Finalmente, en el capítulo 6 se mencionará las conclusiones, los trabajos futuros, limitaciones y los puntos a mejorar del alcance final del trabajo de investigación propuesto.

CAPÍTULO 2: ESTADO DEL ARTE

El presente capítulo tiene como finalidad revisar y exponer los modelos o métodos para la detección de fraude en la facturación de agua, para este fin se ha seguido una metodología de cuatro fases: planificación, ejecución, resultados y análisis.

2.1. Metodología

Para la elaboración del estado del arte se ha tenido en cuenta las siguientes fases: planificación, ejecución y resultados. Esta metdología ha sido usada por (Kitchenham, 2007) y (Ardito, 2015).

- Planificación: Se elaborarán las preguntas de investigación, incluyendo los criterios de inclusión y exlusión para la respectiva selección de los artículos.
- Ejecución: Se aplica el protocolo definido para la filtración de los artículos de acuerdo a los criterios de inclusión y exclusión.
- Resultados: Se presenta los resultados obtenidos en la sección de la ejecución.
- Análisis: Se responderán las preguntas planteadas en la sección de la planificación.

2.2. Planificación

Para lograr el propósito de la investigación de un sistema inteligente para la detección de fraude en facturación de agua, se proponen las siguientes preguntas:

Q1: ¿Qué variables son las que se toman en cuenta en la identificación de anomalías de la facturación de fraude de agua?

Q2: ¿Qué técnicas son las más acertadas para este tipo de fraude?

Q3: ¿Qué herramientas tecnológicas de minería de datos se han usado?

Para este fin se realizaron las búsquedas de artículos científicos o revistas con factor de impacto SJR: ScienceDirect, DOAJ, Taylor and Francis, IEEE Xplore y Elsevier. Así mismo, se usó las siguientes cadenas de búsqueda: “fraud water”, “fraud detection service water”, “billing fraud water” y “fraud water comsutption”, todos pertenecientes al periodo 2010 – 2017. Se usó los criterios de exlusión de validación en los documentos, si se dan el caso en diferentes partes del mundo o el porcentaje de exactitud en las técnicas usadas.

Criterios de inclusión	Criterios de exclusión
• Presentan técnicas o modelos para detección de fraude en consumo de agua. • Presentan el uso de datasets. • Proponen nuevos métodos para la detección de fraude en consumo de agua. • Respondan a las preguntas de investigación.	• Papers donde no usen como solución de aprendizaje para detección de fraude de consumo de agua. • Posters, editoriales.

Tabla 2.1. Criterios de inclusión y exclusión

2.3. Ejecución

La implementación del protocolo se explica en la figura 2.1, en donde se observa que, de los bancos de artículos científicos iniciales, se encontró 82 artículos mediante el criterio de exclusión e inclusión establecidos, string y combinación de palabras, luego por criterio de exclusión se redujeron a 8 artículos y 3 tesis.

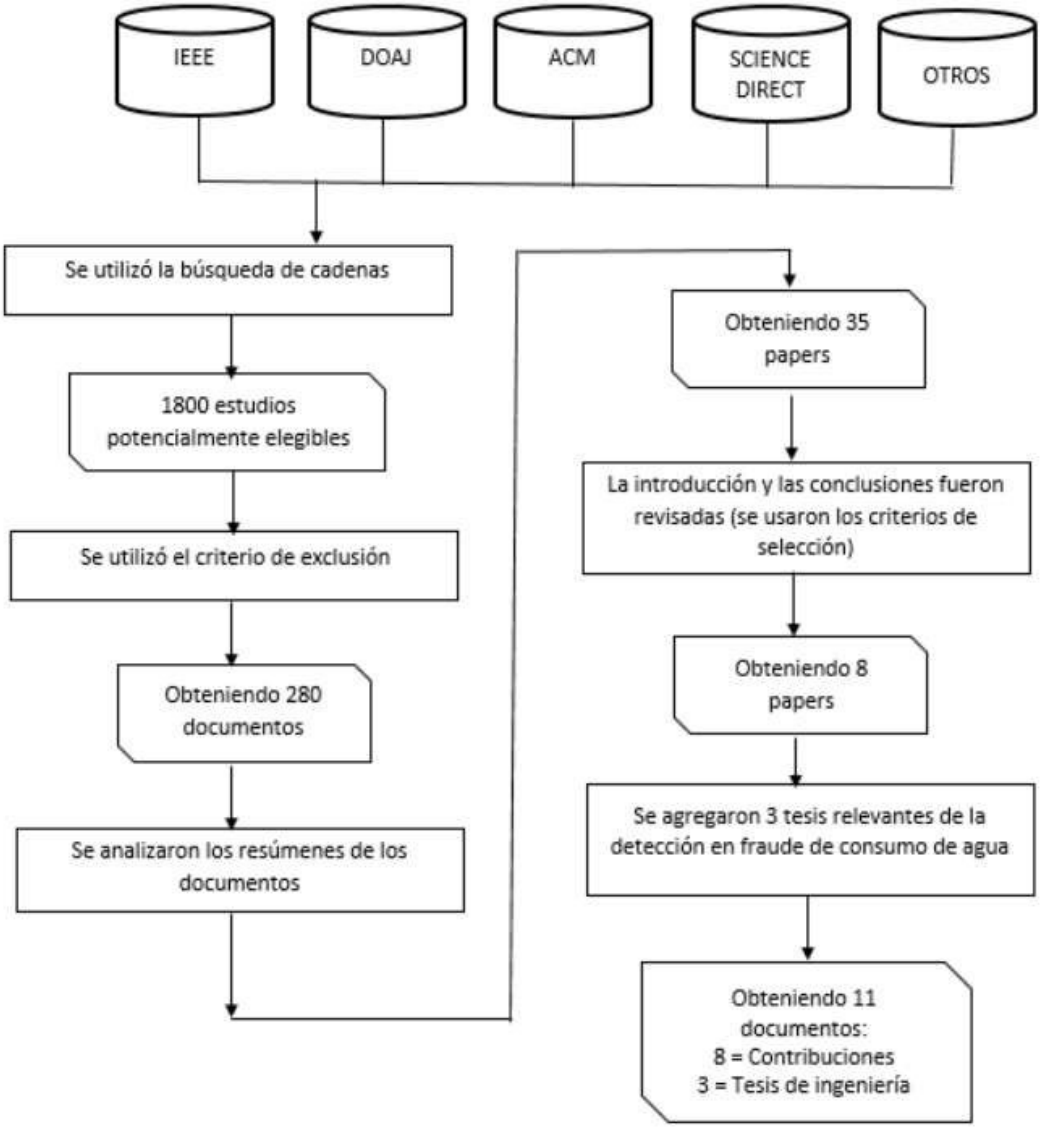


Figura 2.1. Diagrama de flujo de selección de artículos para el estado de arte

2.4. Resultados

2.4.1. Documentos seleccionados

Luego de ejecutar la planificación de acuerdo con la figura 2.1., se muestra los 11 documentos obtenidos en la tabla 2.2.

ID	Título	Autor	Journal (Fuentes)
A01	Reconocimiento de patrones en usos residenciales finales de agua utilizando redes neuronales artificiales y otras técnicas de aprendizaje automático	Ibañez, Días, *et al.* (2017)	Computing and Control for the Water Industry
A02	Regresión vectorial de clústeres y soporte para predicción de demanda de agua y detección de anomalías	Candelieri (2017)	Water
A03	Sistema Inteligente para detectar fraude en el servicio de Agua Potable de una empresa Sanitaria	Palomino y Rivera (2016)	Tesis de ingeniería
A04	Una aproximación a la detección de manipulación de medidores de agua	Monedero, Biscarri, *et al.* (2015)	International Conference on Knowledge Based and Intelligent Information and Engineering Systems
A05	La detección del contador del agua de subregistro, utilizando algoritmos estadísticos	Peña, Roldán, et *al.* (2015)	American Society of Civil Engineers
A06	La determinación de la cantidad y localización de fugas en el suministro de redes de agua utilizando una red neuronal mejorada por el Bat optimización del algoritmo	Faghafur, Reza, *et al.* (2014)	Civil Engineering and Urbanism
A07	Modelo de detección de fraude en clientes del servicio de agua potable de una empresa sanitaria	Victoria, Patiño, *et al.* (2014)	Tesis de ingeniería
A08	Un sistema para la detección del fraude en la compañía abastecimiento de agua	De castro, Cerqueira, *et al.* (2014)	Interciencia
A09	El consumo de agua, detección de fraudes financieros: un modelo basado en las reglas de inducción	Humaid y Barhoum (2013)	Palestinian International Conference on Information and Communication Technology
A10	El Modelo de Detección de Fraude basado en la minería de datos para el sistema de facturación Consumo de agua en MOG	Eyad (2012)	Tesis de ingeniería
A11	Evaluación de las pérdidas aparentes en los sistemas de agua urbanos	Mutikanga, Sharma *et al.* (2010)	Water and Environment

Tabla 2.2. Documentos seleccionados para la detección de fraude en facturación de agua

2.5. Análisis

En esta sección se responderán las preguntas planteadas en la planificación.

2.5.1. ¿Qué variables se consideran para la detección en fraude de facturación de agua?

Del estudio de los documentos seleccionados se han encontrado 38 variables que se usan para la detección de fraude en facturación de agua, el inventario de estas variables es presentado en la tabla 2.3.

Id	Variable	Descripción	Referencia
V01	Identicación de cliente	Identificador para un cliente de servicio de agua o residencia	Eyad (2012), Palomino y Rivera (2016)
V02	Tipo de servicio	El tipo de servicio de facturación (agua o electricidad)	Eyad (2012)
V03	Identicación de medidor	Indica el número del medidor de un cliente	Eyad (2012)
V04	Deuda del cliente	Monto de dinero que adeuda el cliente	Eyad (2012), Palomino y Rivera (2016)
V05	Tipo de calculo	Registro si el consumo automático promedio o cantidad real por el lector	Eyad (2012)
V06	Identicación de ubicación	El número de ubicación para identificar el edificio	Eyad (2012)
V07	Polígono social	Si cliente reside en sector social vulnerable (1) o no (0)	Patiño (2014)
V08	Promedio de consumo de 6 meses	Promedio consumo de m^3 en los últimos 6 meses	Patiño (2014), Palomino y Rivera (2016)
V09	Caída	Magnitud de la caída de consumo de mayor relevancia, en los últimos 48 meses	Patiño (2014)
V10	Irregularidad del consumo	Indica si la última clave manifiesta indicio directo de irregularidad o no	Patiño (2014)
V11	Anomalía del medidor	Indica si la última clave manifiesta indicio de anomalías en el medidor o no	Patiño (2014)
V12	Estado de lectura de medidor	Indica si la última clave correspondió a un proceso de lectura normal o no	Patiño (2014)
V13	Número de claves iregulares de los últimos 12 meses	N.° de claves consideradas “irregulares” en los últimos 12 meses	Patiño (2014)
V14	Número de claves anónalas	N.° de claves consideradas “anómalas” dentro de las últimas 12 claves	Patiño (2014)
V15	Número de claves no efectivas	N.° de claves consideradas “no efectivas” dentro de últimas 12 claves	Patiño (2014)

V16	Número de claves de inmueble deshabitado	N° de claves consideradas “de inmueble deshabitado” dentro de últimas 12 claves	Patiño (2014)
V17	Número de claves normales de los últimos 12 meses	N.° de claves consideradas “normales” dentro de las últimas 12 claves.	Patiño (2014)
V18	Número de claves distintas de los últimos 12 meses	N.° de claves distintas de “N” dentro de las últimas 12 claves	Patiño (2014)
V19	Desviación estándar de bajada	Magnitud de la caída de la desviación estándar de mayor relevancia, en los últimos 48 meses	Patiño (2014)
V20	Desviación estándar de subida	Magnitud del alza de la desviación estándar de mayor relevancia, en los últimos 48 meses	Patiño (2014)
V21	Promedio consumo	Promedio de consumo de últimos 6 meses de vecinos del cliente	Patiño (2014)
V22	Concentración de irregularidades	Concentración de irregulares en el sector en que sitúa el cliente	Patiño (2014)
V23	Desaviación estándar de consumo de 12 meses	Desviación estándar normalizada de los últimos 12 consumos mensuales	Patiño (2014)
V24	Desviación estándar de consumo de 6 meses	Desviación estándar normalizada de los últimos 6 consumos mensuales	Patiño (2014)
V25	Consumo de 12 meses	Porcentaje de consumos nulos en últimos 12 meses.	Patiño (2014),
V26	Valores de lecturas	Últimos 24 valores de lectura del medidor.	Monedero (2015)
V27	Consumo	Los valores de consumo en litros antes de proceso de normalización.	Monedero (2015)
V28	Personas por hogar.	Número de personas que habitan un hogar	Monedero (2015)
V29	Posición de fuga	Indica en que sector donde se reside se encuentra la fuga de agua.	Faghafur (2014)
V30	Cantidad de fuga	Indica la cantidad de agua consumida por el usuario.	Faghafur (2014)
V31	Posición del número mínimo de mediciones de presión	Valor de referencia.	Faghafur (2014)
V32	Cantidad de fuga pronosticada	Indica los puntos que son propensos a una fuga de agua por averías.	Faghafur (2014)
V33	Error medio.	Error medio porcentual en cada prueba.	Faghafur (2014)
V34	Consumo medio	Indica la media entre el consumo actual y los 2 meses anteriores.	De Castro (2014)
V35	Desviación estándar	Valor que servirá de referencia para el análisis final.	De Castro (2014)
V36	Coeficiente de variación	Valor que servirá de referencia para el análisis final.	De Castro (2014)
V37	Número de muestras	Cantidad de muestras tomadas para la prueba.	Multikanga (2010)
V38	Cantidad de conteo de pulsos	Los registros de datos de una semana se llevaron a cabo a través del conteo de pulsos en un intervalo de 10 s.	Multikanga (2010)

Tabla 2.3. Variables usadas para la detección de fraude en la facturación agua

2.5.2. ¿Qué técnicas son las más acertadas para este tipo de fraude?

En la tabla 2.4 se muestra las técnicas aplicadas para la detección de fraude en la facturación del consumo de agua, su nivel de exactitud en la detección, el tamaño de registros (instancia) para la validación y la referencia.

Técnica	**Exactitud**	**Instancias**	**Referencia**
Árbol de Decisión	75%	7252	Patiño (2014)
Redes Neuronales	72,8%	7252	
Regresión Logísitca	71,6%	7252	
Redes neuronales y K-Nearest-Neighbor (Vecino más cercano)	98%	55	De Castro (2014)
Detección de datos atípicos	88,2%	7628	Faghafur (2014)
Anormalmente bajo consumo	80%	720.000	Eyad (2012)
Coefciente de correlación de Person	7%	859	Monedero (2015)
Algorítmos estadísticos	7%	250	Multikanga (2010)
Máquina de Soporte Vectorial	95,7%	4814	Palomino y Rivera (2016)

Tabla 2.4. Tabla de análisis de las técnicas usadas en los documentos de investigación

Observe que presentan mayor exactitud son las de redes neuronales - vecino más cercano (De Catro, 2014), pero la instancia de pruebas es pequeña; y la técnica de datos atípicos (Faghafur, 2014). Sin embargo, la exactitud no puede ser considerada un criterio suficiente para identificar la técnica más acertada, en vista que ella depende de los datos, el tamaño de los datos y el escenario.

2.5.3. ¿Qué herramientas tecnológicas de minería de datos se han usado?

Las herramientas tecnlogícas de minería de datos que se emplearon en la literatura son: Weka 3.8, la cual contiene algoritmos de Machine Learning, Máquinas de Soporte Vectorial (Support Vector Machine) (Palomino y Rivera, 2016), SPSS v 17 , Software estadístico informático muy usado en las ciencias sociales y aplicadas (De Castro, 2015), (Patiño, 2014), Trace Wizard® (Aquacraft) o Identiflow, se basa en árboles de decisión simples para clasificar los eventos de uso del agua, Rapid Miner 5.1, programa informático para el análisis y minería de datos que permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico (Eyad, 2012), (Humain, 2013), EPANET2.0., crea modelos matemáticos de Sistemas de Distribución de Agua Potable de cierta complejidad (Faghafur, 2014)

CAPÍTULO 3: MODELO KDD PARA LA DETECCIÓN DE FRAUDE EN LA FACTURACIÓN DE AGUA

En el presente capítulo identificaremos y justificaremos la metodología más adecuada, el diseño de la solución, la descripción del diseño, el modelo a proponer y, por último, desarrollaremos la propuesta del modelo para la detección de fraude en la facturación de agua potable.

3.1. Justificación de la metodología

Se justifica la metodología KDD como base para el desarrollo de esta investigación debido a que es considerada un proceso metodológico para encontrar un "modelo" válido, útil y entendible que describa patrones de acuerdo a la información, y que ha sido aplicado en diversos problemas de detección de fraude, como Fraud Detection in High Voltage Electricity Consumers Using Data Mining (Detección de fraude en consumidores de electricidad de alto voltaje utilizando minería de datos) (Cabral *et al.*, 2008), Using the rough set theory to detect fraud committed by electricity customers (Utilizando la teoría de conjuntos para detectar el fraude cometido por los clientes de electricidad) (Spiric *et al.*, 2014).

3.2. Modelo KDD propuesto

En la figura 3.1 se presenta las estapas que conforman la metodología propuesta para la detección en fraude de facturación de agua.

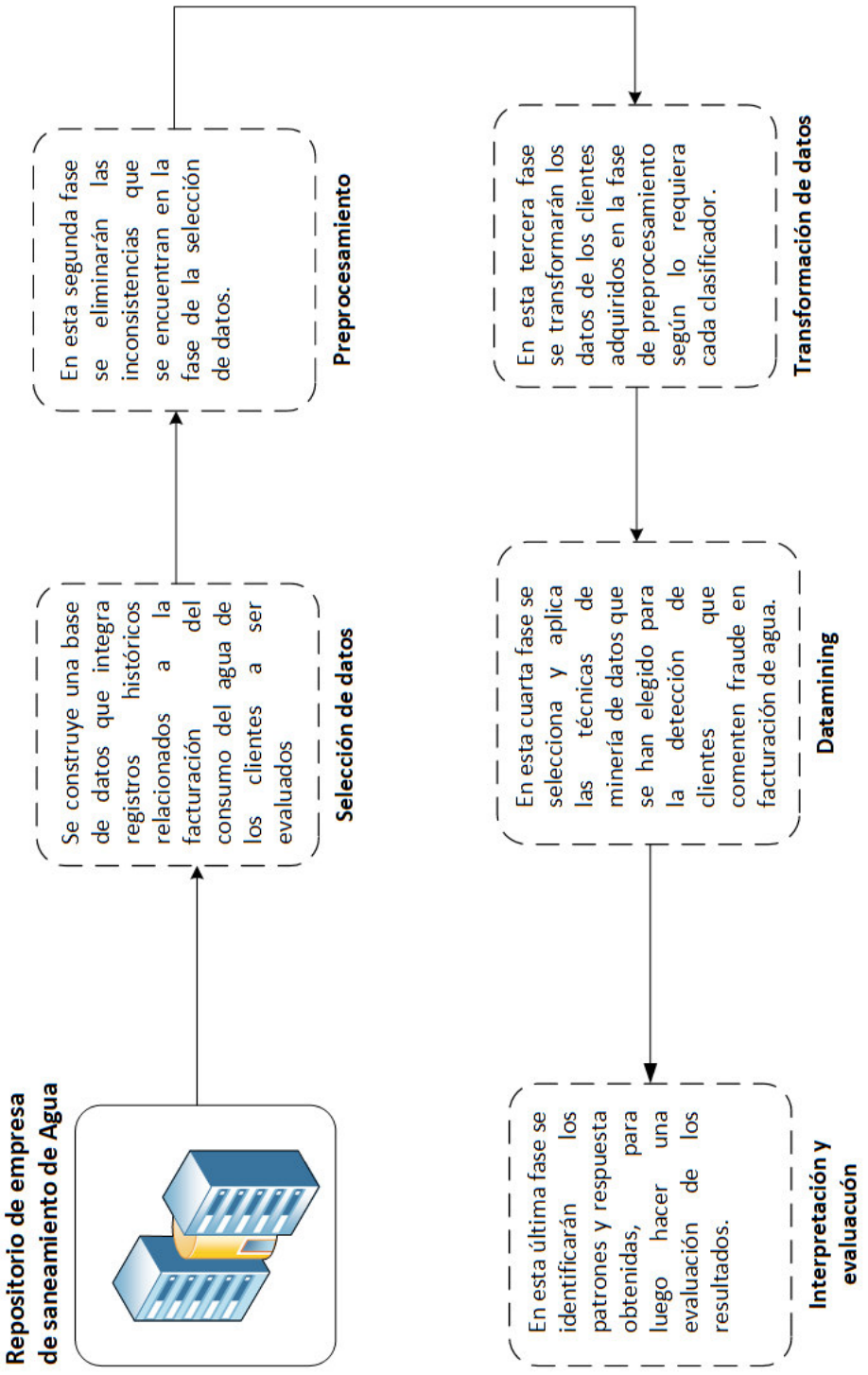


Figura 3.1. Modelo KDD propuesto para la detección de fraude en la facturación de agua

3.3. Descripción del modelo

3.3.1. Selección de datos

En esta fase se construye una base de datos que integra registros históricos relacionados a la facturación del consumo del agua de los clientes a ser evaluados, a partir de diversos repositorios de la organización. Se sugiere que dicha base de datos considere en lo posible los datos asociados a las variables dadas en la literatura (Ver tabla 2.3).

3.3.2. Preprocesamiento de datos

Se tratan los casos duplicados, missings (datos faltantes) y outliers (datos fuera de rango) (Fayyad *et al*, 1996) normalización de datos (Dorian Pyle, 1999).:

- Caso de datos duplicados:
  - Si presenta usuarios repetidos, eliminarlos.
- Caso de Missings:
  - Para el caso de un cliente que posee pocos valores faltantes, se procede a imputar su valor por el promedio.
  - Ignorar la variable si presenta un alto porcentaje de missings.
  - Ignorar la fila si presenta un alto porcentaje de missings.
- Caso de Outliers:
  - Elminar la variable en la base de datos si la variable presenta muchos outliers.
  - Eliminar los registros de un cliente si este presenta muchos outliers en las variables.

- Normalizar datos:

  Normalizar significa que cada dimensión puede construirse de modo que sus valores máximo y mínimo sean los mismos. Es muy conveniente construir el rango de escala de modo que el valor máximo sea 1 y el mínimo 0.

$$v_n = \frac{v_i - \min(v_1 \ldots v_n)}{\max(v_1 \ldots v_n) - \min(v_1 \ldots v_n)}$$

Dónde:

v_n: Valor normalizado.

v_i: valor de instancia.

Esta expresión toma cualquier valor y lo transforma en otro número. Si el valor de entrada está dentro de los límites, la salida estará entre 0 y 1. Cualquier valor fuera de los límites quedará fuera del rango 0–1 (Dorian Pyle, 1999).

3.3.3. Transformación de datos

Se procederá a transformar los datos según los requerimientos de cada modelo de clasificación (si se requiere variables estandarizada, normalizada, escalonada, entre otros) (Fayyad *et al*, 1996)

3.3.4. Data mining

Esta fase se trata del análisis exploratorio y la selección de modelos: elegir el (los) algoritmo (s) de análisis de datos y seleccionar los métodos que se utilizarán para buscar patrones de datos.

Este proceso incluye decidir qué modelos y parámetros podrían ser apropiados (por ejemplo, los modelos de datos categóricos son diferentes a los modelos de vectores sobre

los reales) y hacer coincidir un método particular de extracción de datos con los criterios generales del proceso KDD (por ejemplo, el usuario final podría estar más interesado en comprender el modelo que en sus capacidades predictivas) (Fayyad *et al*, 1996).

3.3.4.1. Variables

En esta etapa se van a saleccionar las variables que influyen en la detección de fraude de facturación de agua. En la literatura existen 38 variables (ver tabla 2.3) que pueden ser usadas para el modelo híbrido predictivo y a su vez consideramos para este modelo las variables que se ven en siguiente tabla 3.1 que, según (Monedero, 2012) (Patiño, 2014) (Eyad+2012), han proporcionado mayor tasa de detección de fraude de facturación de agua.

	Variable	**Descripción**	**Tipo de variable**	**Valores**
Variables consideradas	Identificador de cliente	Identificador del acuerdo, con la cual se puede identificar al cliente.	Numérica	0-10000
	Cantidad de consumo mensual	Consumo mensual de agua (m^3). Esta variable consta de atributos que van desde 03/2000 hasta el 02/2012.	Numérica	0-1000
	Facturas pagadas	Es el número de facturas que han sido pagados totalmente por los clientes	Numérica	0-100
	Pagos pendientes del cliente	Es el porcentaje de cantidad de pagos de los clientes con respecto a la cantidad de visitas para que realice el pago de facturas	Numérica	0-100
	Fraude	Indica si el cliente está categorizado como fraudulento o normal. Con los valores YES/NO	Categórica	YES-NO

Tabla 3.1. Variables consideradas para modelo híbrido

3.3.4.2. Técnicas de Machine Learning

El modelo propuesto está basado en las ténicas de Regresión logística lineal, Árbol de decisión y Máquina de Soporte Vectorial, técnicas que fueron seleccionadas de acuerdo a la literatura que son las que presentan mayores tasas de detección siendo respectivamente 90 % (De Castro, 2014), 75 % (Patiño, 2014) y 92,29% (Eyad, 2012) 95,7 % (Palomino y Rivera, 2016) las cuales obtuvieron también un mayor porcentaje de detección de fraude de facturación de agua en elcapítulo de la validación.

3.3.4.3. Modelo híbrido

Se procederá a construir el modelo híbrido propuesto para la detección de fraude de facturación de agua que consta de: variables, clasificadores de minería de datos y parámetros.

Para la construcción del modelo híbrido se seleccionará una base datos de datos mencionada anterioremente en 3.3.1. Del problema de facturación de fraude de agua. Se realizará diversas pruebas numéricas para identificar los clasificadores que arrojen un mejor resultado. La cantidad de clasificadores elegidos tiene que ser un número impar y mayor que 1, tal como se muestra en el esquema de la figura 3.2.

El modelo híbrido se basa en los siguientes clasificadores y los cuales fueron seleccionados por sus respectivos resultados que se presentarán en el capítulo 5:

- Regresión logistica lineal.
- Árbol de decisión.
- Máquina de Soporte Vectorial.

Estudios han demostrado que un enfoque híbrido de tres clasificadores que utilizan la técnica de votación para un mismo conjunto de datos ha superado a estos clasificadores por separado (Meka y Amit, 2014).

De los cuales presentan resultados exitosos en diversos problemas de Data Mining (los cuales arrojaron mayor porcentaje de exactitud), sencillos en interpretación de los resultados y con los cuales, usando el criterio de la técnica de votación, donde la mayoría de resultados arrojados de cada uno de los clasificadores, dice que si se obtiene un resultado negativo éste está cometiendo fraude. La base de datos de prueba usada por cada uno de los clasificadores y aplicando la técnica de votación mayoritaria, obtendrá el resultado final, tal como se aprecia en la figura 3.2.

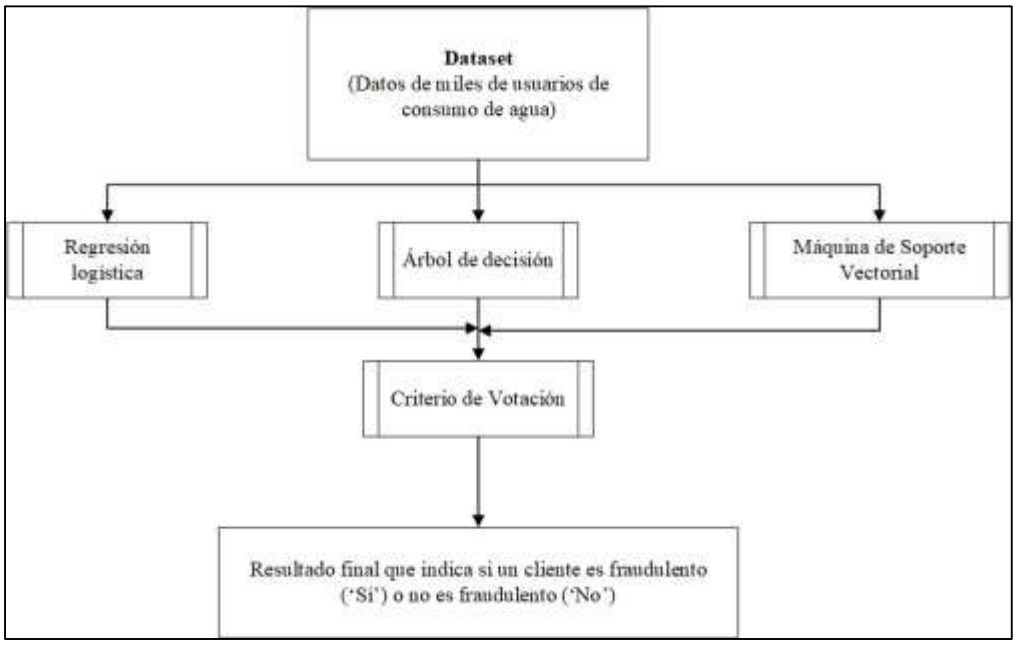


Figura 3.2. Modelo predictivo

En la anterior figura 3.2. se visauliza que mediante el criterio de votación (Meka y Amit, 2014), se decidirá por la respuesta más votada, es decir, si el cliente es detectado como fraudulento o no.

3.3.4.4. Parámetros

En esta fase se van a calibrar los parámetros necesarios de las técnicas de minería de datos que han sido seleccionados. En la tabla 3.2 se mencionarán los parámetros seleccionados (S. C. Chen, 2016) para calibrar el método de Regresión Logística, en la tabla 3.3 se mencionarán los parámetros seleccionados por (Patiño, 2014) para calibrar el método de Árbol de decisión. En la tabla 3.4 se mostrarán los parámetros seleccionados por (Palomino y Rivera, 2016) para el método de Máquina de Soporte Vectorial.

Regresión Logísitca						
Variable	B	E.T.	Wald	gl	Sig	Exp(B)
Valor	0,003	0,009	0.128	1	0.721	1,003

Tabla 3.2. Parámetros de Regresión Logística

Árbol de Decisión			
Variable	Número de Nodos	Número de pliegues poda	Proporción de registros
Valor	3 (Por defecto)	14	0.7

Tabla 3.3. Parámetros de Árbol de decisión

Máquina de Soporte Vectorial		
Variable	Costo	Gamma
Valor	30	1

Tabla 3.4. Parámetros de Máquina de Soporte Vectorial

3.3.5. Interpretación y evaluación

En esta fase se evaluará los modelos, de forma analítica, obtenidos en la etapa de Data Mining. Para ello, se usará la siguiente matriz, la matriz de confusión (Qasem y Mahmoud, 2018) mostrada en la Tabla 3.5:

Matriz de Confusión		**PREDICCIÓN**	
		NI	I
REAL	NI	A	B
	I	C	D

Tabla 3.5. Matriz de Confusión para evaluar el modelo (Qasem y Mahmoud, 2018)

- NI: clientes no irregulares (normales).
- I: clientes irregulares.
- A: número de clientes clasificados por el modelo como "normales" y que en la realidad lo son.
- B: número de clientes clasificados por el modelo como "irregulares", pero que en la realidad son normales.
- C: número de clientes clasificados por el modelo como "normales", pero que en la realidad son irregulares.
- D: número de clientes clasificados por el modelo como "irregulares" y que en la realidad son irregulares.
- Con los indicadores de desempeño:
  - ✓ $Eficacia = \frac{D}{C+D}$
  - ✓ $Eficiencia = \frac{D}{B+D}$

La eficacia quiere decir que porcentaje del total de clientes verdaderamente irregulares se clasificaron correctamente como irregulares, la eficiencia, por otro lado, se define cuántos de los clientes predecidos como irregulares por el modelo, en verdad lo son (Figueroa, 2009).

Se interpreta las respuestas luego de aplicar la técnica de votación a los resultados de los clasificadores, para ello se considera los tipos de errores usados para el tema de detección de fraudes (Ellas, 2012) (Figueroa, 2009):

Error Tipo I: Cuando el modelo detecta a un cliente irregular como normal.

Error Tipo II: Cuando el modelo detecta a cliente normal como irregular.

CAPÍTULO 4: DESARROLLO DEL SISTEMA INTELIGENTE PARA LA DETECCIÓN DE FRAUDE EN EL SERVICIO DE AGUA POTABLE

En el presente capítulo se describirá como se ha encapsulado la herramienta weka en una aplicación web, donde el usuario tendrá un entorno amigable del sistema y sencillo de usar. A continuación, se describirá su arquitectura, la funcionalidad del sistema y sus respectivos roles.

4.1. Arquitectura del Sistema

De acuerdo a los requerimientos funcionales que se captaron del sistema, este será desplegado en un entorno vía web, con una arquitectura Cliente-Servidor, en donde se usó en el lado del backend el lenguaje de programación Python con el framework Django y para el lado del cliente Angular 6, PostgreSQL para la base de datos, y la librería weka para Python que nos permite la creación de los modelos inteligentes para detectar el fraude. La arquitectura se muestra en la siguiente figura 4.1.

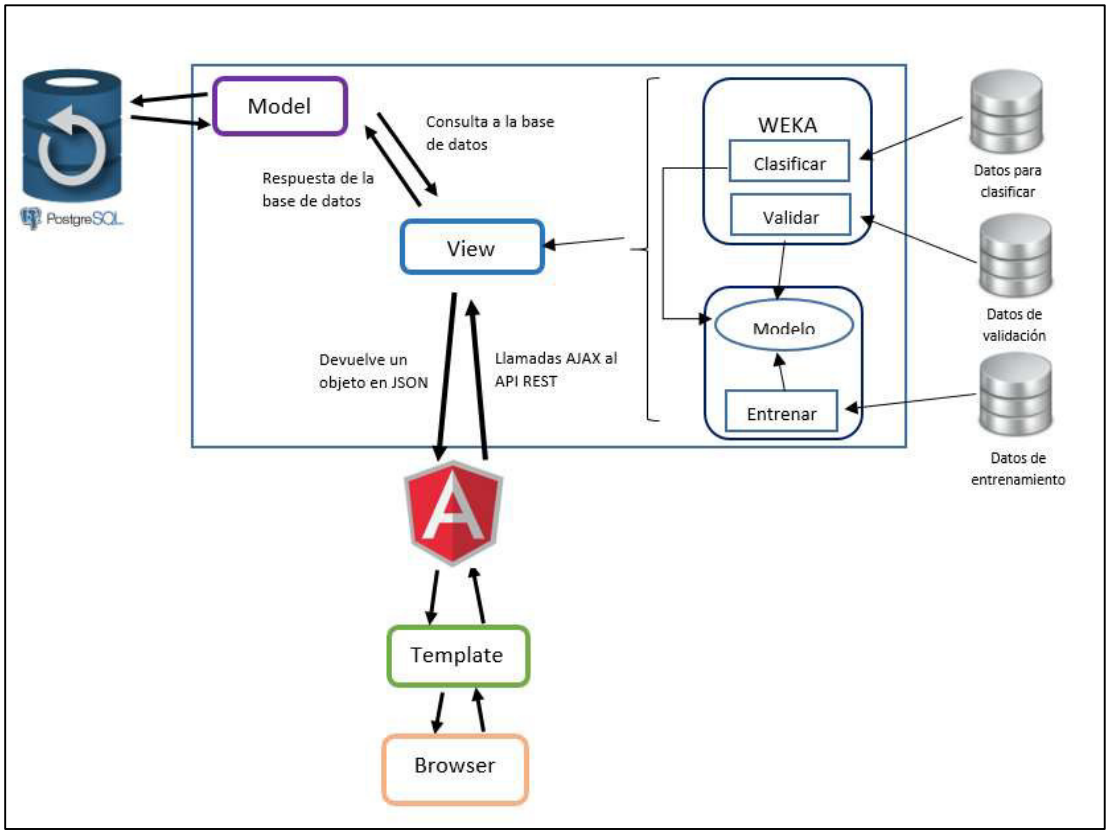


Figura 4.1. Arquitectura del Sistema Web

La característica principal de esta arquitectura es que el mismo sistema solicita sus propios recursos y el servidor responde inmediatamente a sus solcitudes, esto quiere decir que no requiere de otras aplicaciones para proporcionar parte del servicio.

La arquitectura para el sistema está basada en el patron de diseño MVC (Modelo, Vista, Controlador), en este patrón estas 3 capas, la lógica de acceso a la base de datos, la lógica de negocios y la lógica de presentación, pero para este caso el framework Django basado en Python adquiere un patrón de diseño MTV (Model, Template y View) 'Model', acceso

a la base de datos, ‘Template’, la capa de presentación, qué datos son mostrados o no en una web y ‘View’, la capa lógica de negocios que contiene la lógica que accede al modelo.

4.1.1. Descripción del Sistema

Se desarrollará un sistema que tendrá como objetivo principal la detección de clientes fraudulentos y como puntos secundarios el entrenamiento del modelo híbrido e ingreso de parámetros para las técnicas seleccionadas, respectivamente.

El alcance abarca lo siguiente:

- ✓ Crear un sistema web que valide el ingreso al sistema.
- ✓ Permitir la creación de usuarios y asignación de permisos de acuerdo a los diferentes roles.
- ✓ La carga de los datos en formato CSV (archivo que es delimitado por comas) que servirá para el entrenamiento del modelo y validación de datos del modelo final.
- ✓ Configuración de parámetros para cada uno de los clasificadores desde el sistema web.
- ✓ Generar reportes de clientes que cometen fraude en su facturación de agua.

4.2. Modelado del Sistema

Para el modelado del sistema, se describe a los usuarios del sistema, los casos de uso respectivos y la presentación de las funcionalidades a través de las interfaces de usuario que se han desarrollado. En la siguiente figura 4.2, se muestra la interfaz de ingreso para el sistema.

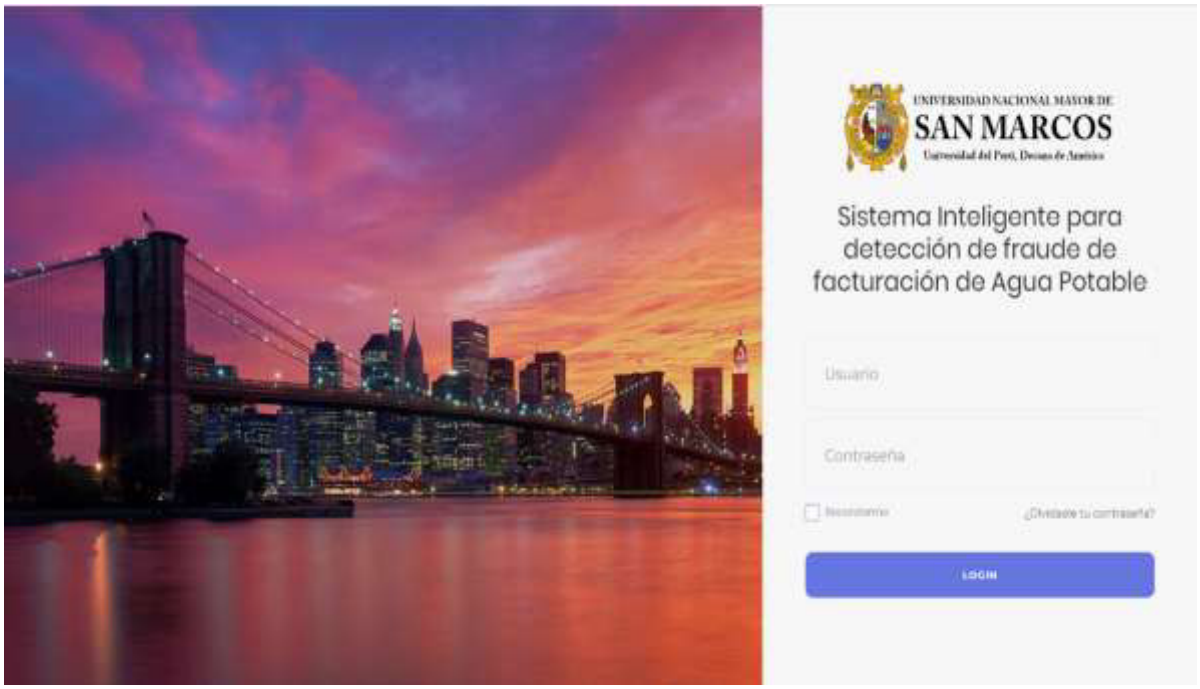


Figura 4.2. Interfaz de ingreso del Sistema de detección de fraudes

4.2.1. Usuarios del Sistema

El sistema tendrá tres tipos de roles para los usuarios correspondientes.

- Usuario: Administrador

Descripción: Usuario encargado del preprocesamiento de los datos de entrenamiento del modelo híbrido y actualización del modelo existente.

Funciones:

- Entrenar los modelos seleccionados para la correcta detección de clientes fraudulentos es de vital importancia, puesto que cada cierto tiempo el comportamiento del consumo de los clientes variarán. También podemos variar ciertos parámetros de cada una de las técnicas de clasificación y observar la tasa de

exactitud en la detección de fraude, tal como se puede apreciar en la siguiente figura 4.3:

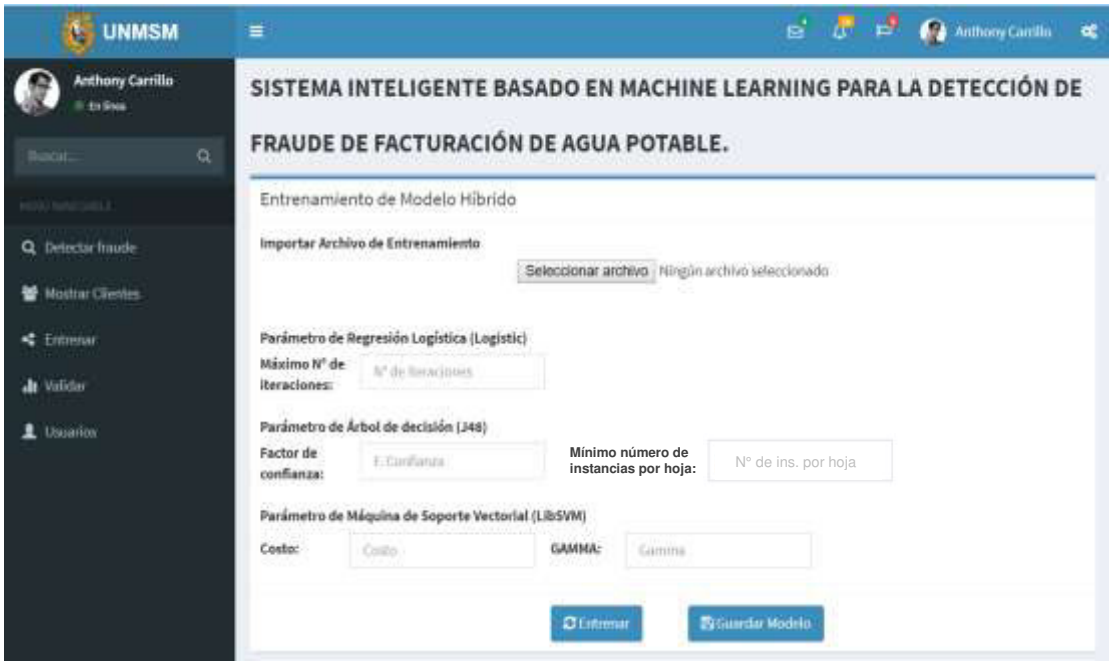


Figura 4.3. Entrenamiento del modelo híbrido

- Configurar y validar cada uno de los clasificadores seleccionados, es aquí donde el usuario ingresa los datos de aquellos clientes que no hayan sido utilizados para el entrenamiento previo. Así mismo, se puede observar la tasa de exactitud de la validación, tal como se aprecia en la figura 4.4.

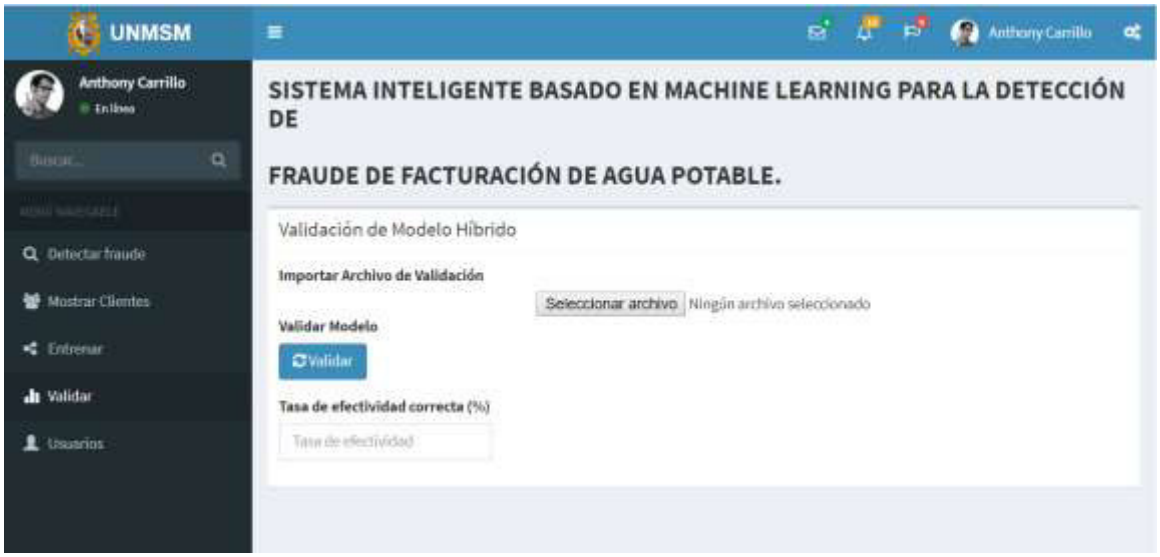


Figura 4.4. Validación del modelo híbrido

➢ Usuario: Gerencia

Descripción: Usuario encargado de detectar fraude y ver los reportes de clientes fraudulentos.

Funciones:

- Identificar el fraude mediante la carga de clientes de los cuales se desconoce fraude alguno o no en el sistema, usará el modelo híbrido entrenado y luego validado, y finalmente mostrará a los clientes que estén cometiendo fraude.

Figura 4.5. Interfaz de detección de fraude

- Generar reportes, luego de detectar fraude en los clientes, el usuario puede obtener un listado de los clientes fraudulentos que mostrará el id del cliente y la opción de ver detalles estadísticos.

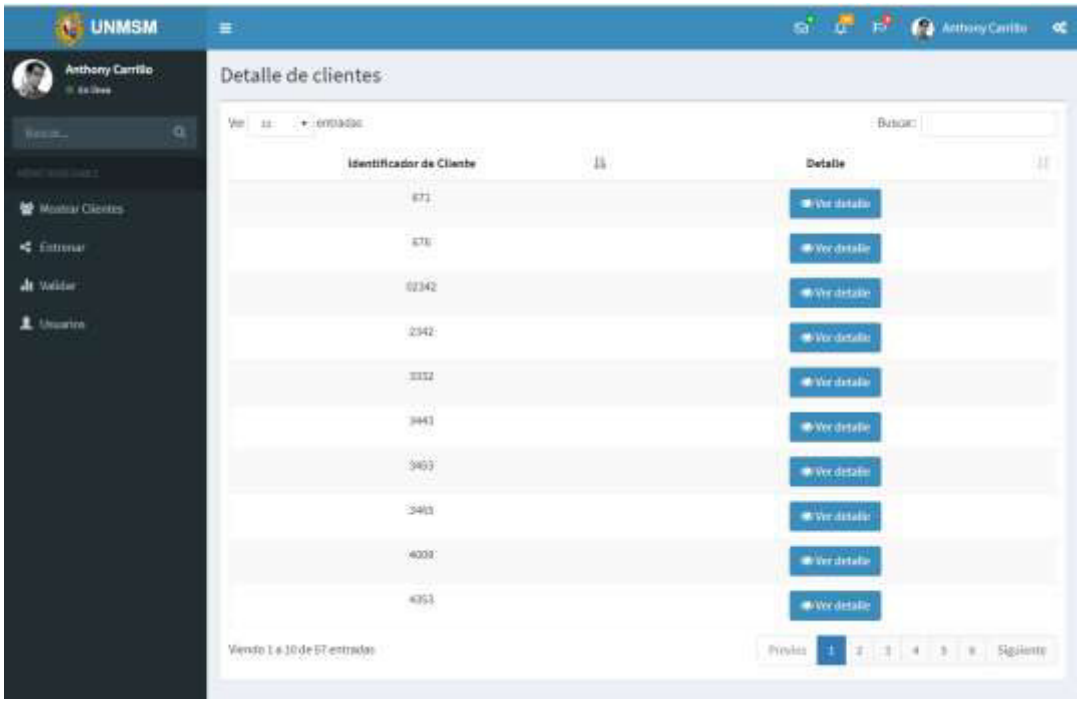


Figura 4.6. Reporte de lista clientes detectados

4.2.2. Casos de uso del Sistema

Ya conocidos los usuarios del nuestro sistema en mención, se definirá los requerimientos haciendo uso de la técnica de casos de uso. Los casos de uso se han agrupado en paquetes correctamente para poder describir las distintas funcionalidades del sistema. A continuación, en la figura 4.7 se muestra los paquetes que tiene el sistema.

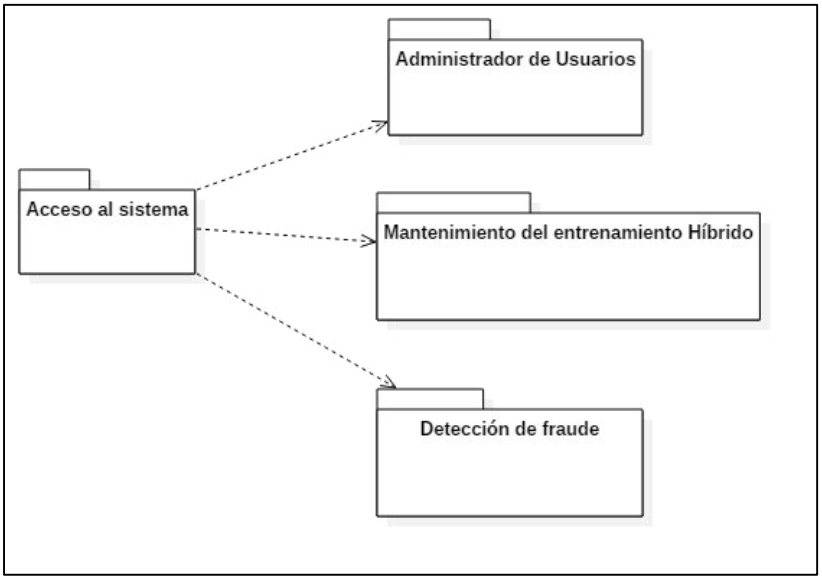


Figura 4.7. Paquetes del Sistema

4.2.1.1. Acceso al Sistema

- CUS: Validar Usuario
  - Descripción: Este caso de uso permite validar el acceso al sistema con los permisos respectivos otorgados.
  - Actores: Jefe de fraude, gerencia, administrador del sistema
  - Flujo de eventos

Flujo básico

1) Empieza cuando el usuario se encuentra en la pantalla principal de acceso al sistema.

2) El sistema mostrará una pantalla de validación de acceso al sistema, con los campos de usuario y contraseña.

3) Ingresar el usuario y contraseña válidas, y dar clic en el botón “Iniciar Sesión”.

4) El sistema comprobará si el usuario existe en la base de datos y si existe, mostrará la interfaz según el rol que tenga el usuario.

5) El caso de uso termina.

El diagrama de actividades del CUS Validar Usuario se muestra en la siguiente figura 4.8:

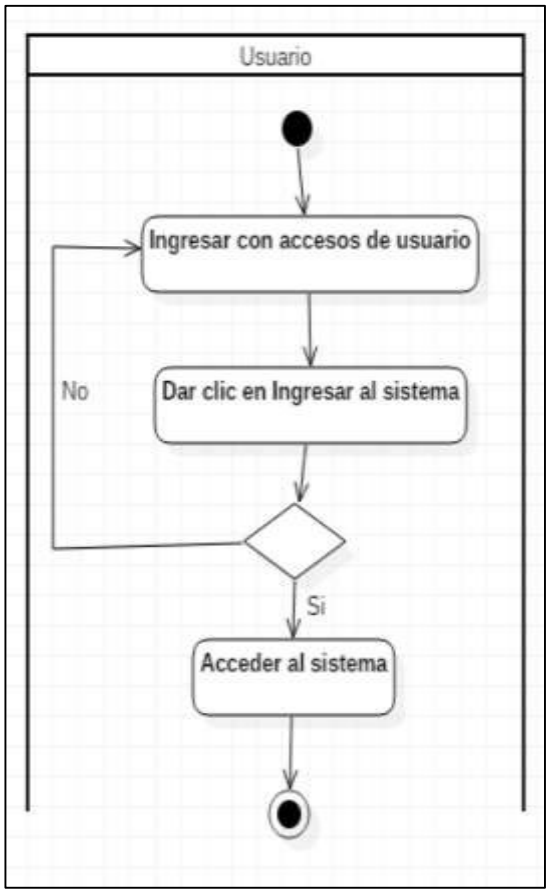


Figura 4.8. Diagrama de actividades del caso de uso Validar Usuario

- Precondiciones: Ninguno
- Poscondiciones: El usuario al ser validado ingresa al sistema.
- Diagrama de caso de uso: El diagrama se presenta en la siguiente figura 4.9:

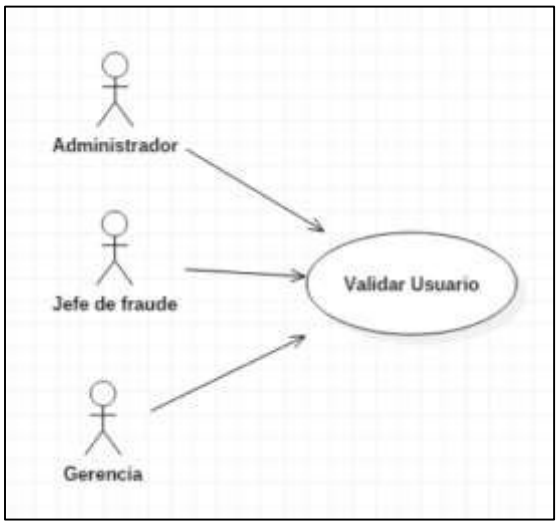


Figura 4.9. Diagrama de CUS Validar Usuario

4.2.2.2. Detección de Fraude

➢ CUS: Importar Archivo

- Descripción: Este caso de uso sirve para importar un archivo, tanto para el entrenamiento y validación del modelo o detectar a los clientes que comenten fraude.
- Actores: Jefe de fraude, Gerencia
- Flujo de eventos

- Empieza cuando el usuario le da clic en el botón de "Importar Archivo".
- El sistema muestra una ventana del explorador del sistema operativo para seleccionar el archivo.
- El usuario selecciona el archivo en formato CSV y da clic en seleccionar, luego el usuario da clic en el botón "Cargar Archivo".
- El sistema carga el archivo y muestra el mensaje de "cargó correctamente el archivo".
- El caso de uso finaliza.

El diagrama de actividades del CUS se muestra en la siguiente figura 4.10:

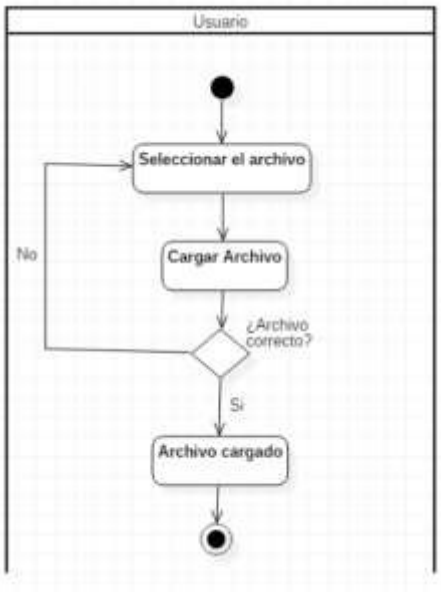


Figura 4.10. Diagrama de Actividades del CUS Importar Archivo

- Precondiciones: El usuario ha sido validado por el sistema.
- Poscondiciones: El usuario importó de manera satisfactoria el archivo en formato CSV.
- Diagrama: El diagrama del CUS se muestra en la siguiente figura 4.11:

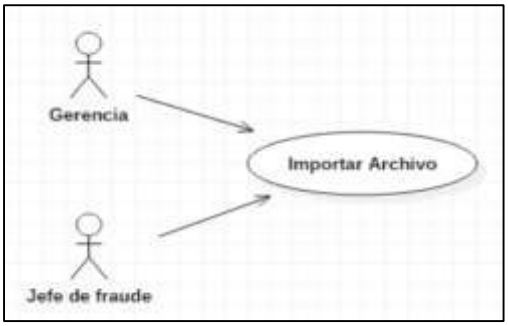


Figura 4.11. Diagrama del CUS Importar Archivo

- CUS: Entrenar modelo

  - Descripción: Este caso de uso permite entrenar el modelo híbrido para luego validarlo.

  - Actores: Jefe de fraude

  - Flujo de eventos

    Flujo básico

    1) El caso de uso comienza cuando el usuario ingresa a la funcionalidad de "Entrenar".

    2) El sistema muestra una interfaz de usuario donde se verá lo siguiente: Importar archivo, parámetros de Regresión Logística, Parámetros de Árbol de decisión, Parámetros de Máquina de Soporte Vectorial, Entrenar y Guardar Modelo.

    3) El usuario da clic en el botón de "Importar Archivo" y seleccionar el archivo de datos.

    4) El usuario ingresa los parámetros de Regresión Logística.

    5) El usuario ingresa los parámetros de Árbol de decisión.

    6) El usuario ingresa los parámetros de Máquina de Soporte Vectorial.

    7) El sistema muestra la opción de "Entrenar Modelo" y "Guardar Modelo".

    8) El usuario da clic en "Entrenar Modelo" y con ello empieza el entrenamiento de los modelos de Regresión Logística, Árbol de decisión y Máquina de Soporte Vectorial, e inmediatamente el entrenamiento del modelo hibrido clasificador por Votación que es basado en los resultados de los anteriores modelos.

9) El sistema entrena el modelo y muestra el porcentaje de precisón en la detección de fraude.

10) El usuario da clic en “Guardar Modelo” y se descargará el modelo híbrido entrenado.

11) El sistema descarga el modelo y muestra el mensaje “Modelo descargado exitosamente”.

12) El caso de uso finaliza.

Flujo alternativo

- En el paso 8, el usuario da clic en la opción de “Entrenar Modelo”. Si el usuario no llegó a importar ni un archivo, el sistema arrojará un mensaje “Debe importar archivo para entrenamiento”, luego continúa el paso 2.
- En el paso 8, el usuario ingresa al módulo de “Entrenar”. Si falto llenar algún parámetro El sistema muestra un mensaje “Llenar todos los campos”, luego continúa el paso 2.

El diagrama de actividades del CUS “Entrenar Modelo” se muestra en la siguiente figura 4.12:

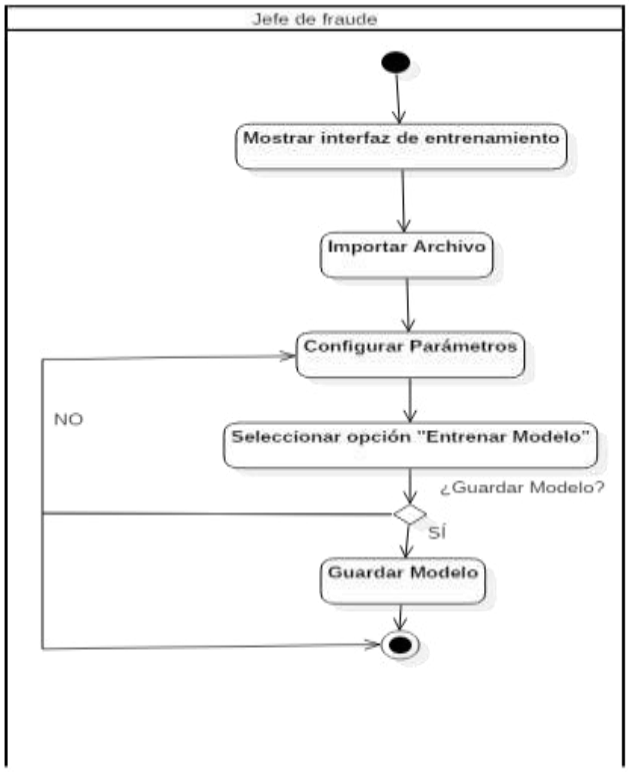


Figura 4.12. Diagrama de actividades del CUS Entrenar Modelo

- Precondiciones: El usuario ha ingresado correctamente al sistema.

- Poscondiciones: El usuario entrenó, validó y guardó el modelo.

- Diagrama: El diagrama del CUS se muestra en la siguiente figura 4.13:

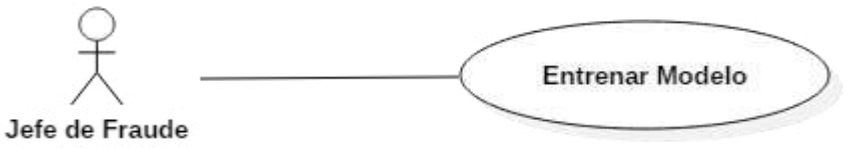


Figura 4.13. Diagrama del CUS Entrenar Modelo

- CUS: Validar Modelo

  - Descripción: En este caso de uso nos permite validar el modelo que se entrenó con nuevos registros.
  - Actores: Jefe de fraude
  - Flujo de eventos

  Flujo básico

  1) El caso de uso empieza cuando el usuario ingresa a la opción de "Validar Modelo"
  2) El sistema muestra en la interfaz de validación, las opciones de "Importar Archivo" y "Validar Modelo".
  3) El usuario hace clic en "Importar Archivo" y seleccionar el archivo de datos para la validación del modelo.
  4) El usuario hace clic en "Validar Modelo".
  5) El sistema carga el modelo híbrido entrenado, realiza la validación y muestra al usuario el porcentaje de tasa de exactitud.
  6) El caso de uso finaliza.

  Flujo alternativo

  - En el paso 5, si existe algún tipo de error en la carga del modelo híbrido, el sistema muestra un mensaje "Error al cargar el modelo".

El diagrama de actividades del CUS Validar Modelo se muestra en la siguiente figura 4.14:

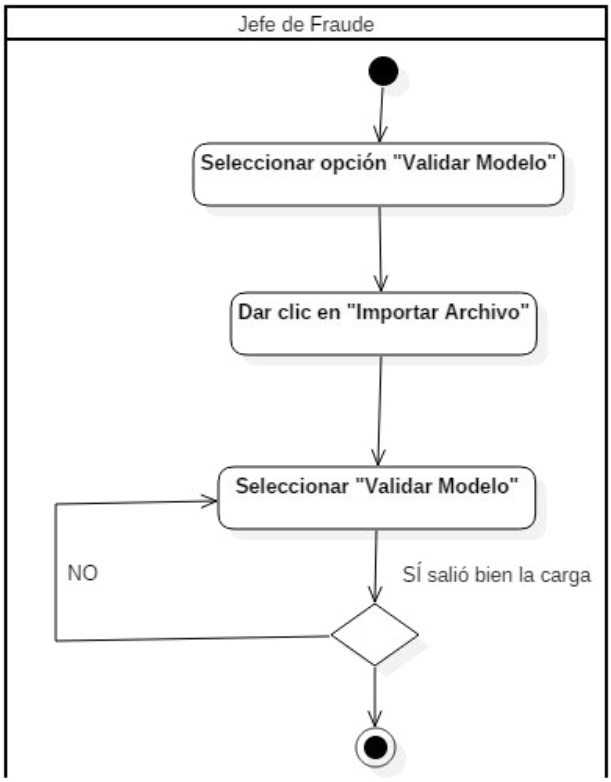


Figura 4.14. Diagrama de actividades del CUS Validar Modelo

- **Precondiciones:** El usuario ha ingresado correctamente al sistema.

- **Poscondiciones:** El usuario validó exitosamente el modelo.

- **Diagrama:** El diagrama del CUS se muestra en la siguiente figura 4.15:

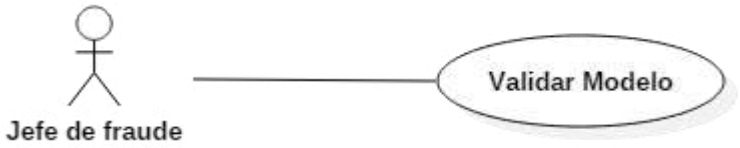


Figura 4.15. Diagrama del CUS Validar Modelo

- CUS: Detectar fraude

  - Descripción: Este caso de uso permite identificar qué clientes son los fraudulentos en su consumo de agua.

  - Actores: Jefe de fraude, Gerencia

  - Flujo de eventos

  Flujo básico

  1) El caso de uso empieza cuando el usuario ingresa a la opción de "Detectar Fraude".
  2) El sistema empieza mostrando en la interfaz, la opción de "Seleccionar archivo".
  3) El sistema muestra en la interfaz la opción "Detectar fraude".
  4) El usuario da clic en "Detectar fraude".
  5) El sistema elige el modelo entrenado (modelo híbrido basado en la combinación de modelos de minería de datos mencionados en el CUS "Entrenar modelo"), luego pasa a ser validado con el archivo importado por el usuario, luego almacena los clientes que han sido detectados como fraudulentos y muestra un mensaje "Nuevos clientes fraudulentos detectados".
  6) El usuario ingresa a la funcionalidad de "Mostrar clientes".
  7) El usuario empieza el caso de uso "Generar detalle de Clientes".

  Flujo alternativo

  - En el paso 5, si no se encuentra ningún cliente que cometa fraude, se muestra el mensaje "No existe clientes frauduelentos".

El diagrama de actividades del CUS se muestra en la siguiente figura 4.16:

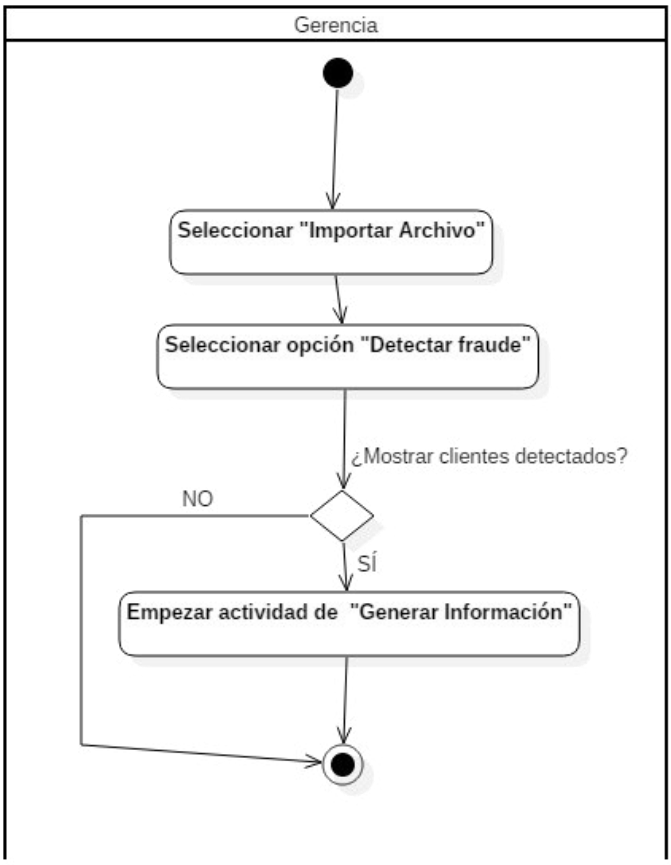


Figura 4.16. Diagrama de actividades del CUS Detectar fraude

- **Precondiciones:** El usuario ha ingresado correctamente al sistema.
- **Poscondiciones:** El usuario visualiza la información de los clientes que han sido detectados como fraudulentos por el sistema.
- **Diagramas:** El diagrama del CUS Detectar fraude, se muestra en la siguiente figura 4.17:

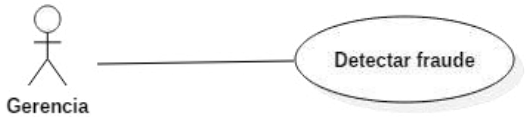


Figura 4.17. Diagrama del CUS Detectar Fraude

- CUS: Generar detalle de clientes
  - Descripción: En este CUS, su respectiva finalidad es la de mostrar información numérica de los clientes que se les detectó como fraudulentos en su consumo de agua.
  - Actores: Gerencia
  - Flujo de eventos

Flujo básico

1) El caso de uso empieza cuando el usario ingresa a la opción de “Mostrar clientes”.
2) El sistema arrojará un reporte donde mostrará una lista de clientes que han sido detectados como fraudulentos, por cada cliente hay una opción de “Ver” para visualizar una data histórica y estadística.
3) El usuario ingresa a la opción de “Información estadística” para un cliente.
4) El sistema muestra la información estadísitca y su respectiva gráfica que describen el comportamiento de consumo de agua mensualmente.
5) El caso de uso finaliza.

El diagrama de actividades del CUS Generar lista de clientes fraudulentos, se muestra en la siguiente figura 4.18:

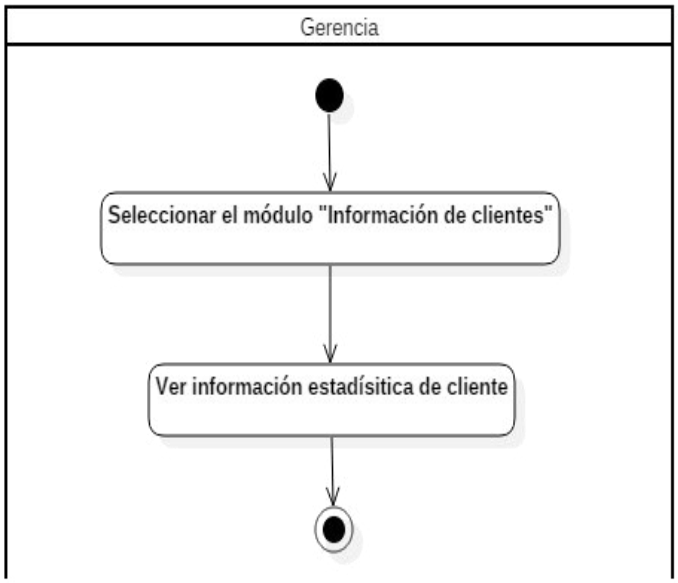


Figura 4.18. Diagrama de actividades del CUS Generar detalle de clientes

- Precondiciones: El usuario ha ingresado al sistema correctamente.
- Poscondiciones: El usuario obtiene información estadística por cliente detectado como fraudulento.
- Diagramas: El diagrama del CUS se muestra en la siguiente figura 4.19:

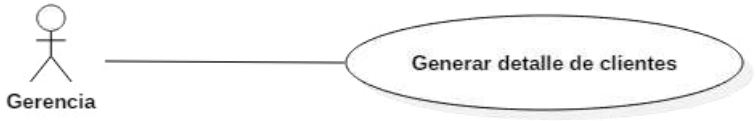


Figura 4.19. Diagrama CUS Generar detalle de clientes

El diagrama y resumen para los casos de uso, respectivamente, se encuentran en el ANEXO 1.

4.3. Conexión Python con Weka

Para realizar un uso satisfactorio de las funcionalidades de weka con Python, se tuvo que importar los siguientes Packages:

- ✓ Python-weka-wrapper
- ✓ Numpy
- ✓ Matplotlib
- ✓ Javabridge

Entre las funcionalidades que la librería weka ofrece está el poder realizar la carga de un archivo en formato CSV, realizar el preprocesamiento, tanto la eliminación de atributos que no ayuden a crear el modelo u otros filtros, así como el entrenamiento y la validación del modelo híbrido. A continuación, se muestra el uso de la librería weka en el sistema en la siguiente figura 4.20:

```
# imports
import weka.core.jvm as jvm
import weka.core.converters as conv
from weka.classifiers import Evaluation, Classifier
from weka.core.classes import Random
import os

# start jvm
jvm.start(packages=True)

# load data
data = conv.load_any_file(os.environ.get("MOOC_DATA")+os.sep+"titanic.arff")
data.class_is_last()

# configure classifier
cls = Classifier(classname="weka.classifiers.trees.J48", options=["-C", "0.3"])

# cross-validate classifier
evl = Evaluation(data)
evl.crossvalidate_model(cls, data, 10, Random(1))

# output statistics
print(evl.summary("=== J48 on anneal (stats) ===", False))
print(evl.matrix("=== J48 on anneal (confusion matrix) ==="))

# stop jvm
jvm.stop()
```

Figura 4.20. Uso de la librería Weka con Python

4.4. Validación del Sistema

4.4.1. Validación de la funcionalidad

Para ver la correcta funcionalidad de nuestro sistema, realizamos una validación a través de la carga de datos de clientes que pueden estar cometiendo fraude.

- Inputs:

Los inputs para el caso de prueba será la data histórica de consumo de clientes durante 12 años, un archivo en formato .CSV, consta de 150 variables y que conforman un registro. La descripción de todas las variables se presentan en el anexo B. En la figura 4.21 se muestran el registro para 4 clientes.

No.	1: AGREEMENT_ID	2: C032000	3: C042000	4: C052000	5: C062000	6: C072000	7: C082000	8: C092000	9: C102000	10: C112000	11: C122000
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
1	10235.0	52.0	52.0	52.0	52.0	63.0	63.0	71.5	71.5	15.5	15.5
2	10282.0	22.5	22.5	24.5	24.5	22.0	22.0	48.5	48.5	13.5	13.5
3	10312.0	26.0	26.0	18.5	18.5	21.0	21.0	36.5	36.5	12.0	12.0
4	13549.0	59.0	59.0	75.0	75.0	80.5	80.5	96.0	96.0	80.0	80.0

12: C012001	13: C022001	14: C032001	15: C042001	16: C052001	17: C062001	18: C072001	19: C082001	20: C092001	21: C102001	22: C112001	23: C122001
Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
52.5	52.5	54.0	54.0	35.0	35.0	42.0	42.0	50.5	50.5	58.0	58.0
25.0	25.0	32.0	32.0	25.0	25.0	34.0	34.0	40.0	40.0	26.5	26.5
22.5	22.5	35.0	35.0	25.5	25.5	34.0	34.0	29.0	29.0	27.0	27.0
61.0	61.0	66.5	66.5	78.5	78.5	86.5	86.5	74.5	74.5	56.5	56.5

24: C012002	25: C022002	26: C032002	27: C042002	28: C052002	29: C062002	30: C072002	31: C082002	32: C092002	33: C102002	34: C112002	35: C122002
Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
49.5	49.5	58.5	58.5	52.0	52.0	51.5	51.5	84.5	84.5	74.5	74.5
16.5	16.5	24.0	24.0	28.5	28.5	26.0	26.0	30.5	30.5	21.5	21.5
24.0	24.0	30.0	30.0	34.0	34.0	44.0	44.0	65.0	65.0	41.5	41.5
60.0	60.0	59.0	59.0	59.0	59.0	69.5	69.5	72.5	72.5	68.0	68.0

36: C012003	37: C022003	38: C032003	39: C042003	40: C052003	41: C062003	42: C072003	43: C082003	44: C092003	45: C102003	46: C112003	47: C122003
Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
70.0	70.0	55.5	55.5	55.0	55.0	55.0	55.0	56.0	56.0	54.0	54.0
21.5	21.5	19.0	19.0	33.0	33.0	53.5	53.5	30.5	30.5	31.5	31.5
2.0	2.0	28.0	28.0	0.0	0.0	27.0	27.0	27.0	27.0	22.5	22.5
70.5	70.5	40.0	40.0	77.5	77.5	80.5	80.5	70.0	70.0	56.0	56.0

48: C012004	49: C022004	50: C032004	51: C042004	52: C052004	53: C062004	54: C072004	55: C082004	56: C092004	57: C102004	58: C112004	59: C122004
Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
37.0	37.0	37.0	37.0	34.5	34.5	60.0	60.0	45.5	45.5	54.5	54.5
26.0	26.0	25.0	25.0	21.5	21.5	26.0	26.0	33.5	33.5	32.5	32.5
27.5	27.5	38.0	38.0	27.5	27.5	33.5	33.5	49.5	49.5	47.0	47.0
68.5	68.5	68.5	68.5	68.5	68.5	81.5	81.5	76.5	76.5	72.5	72.5

60: C012005	61: C022005	62: C032005	63: C042005	64: C052005	65: C062005	66: C072005	67: C082005	68: C092005	69: C102005	70: C112005	71: C122005
Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
29.0	29.0	41.0	41.0	50.0	50.0	53.0	53.0	59.0	59.0	56.0	56.0
17.5	17.5	18.0	18.0	29.0	29.0	27.0	27.0	42.5	42.5	27.5	27.5
37.0	37.0	40.0	40.0	47.5	47.5	47.5	47.5	58.0	58.0	60.5	60.5
62.0	62.0	55.5	55.5	69.0	69.0	61.5	61.5	75.5	75.5	61.5	61.5

84: C012007	85: C022007	86: C032007	87: C042007	88: C052007	89: C062007	90: C072007	91: C082007	92: C092007	93: C102007	94: C112007	95: C122007
Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
52.0	52.0	46.0	46.0	67.0	67.0	53.0	53.0	52.0	52.0	52.0	52.0
27.5	27.5	26.5	26.5	46.0	46.0	30.5	30.5	27.5	27.5	27.5	27.5
50.5	50.5	59.0	59.0	81.5	81.5	54.5	54.5	50.5	50.5	50.5	50.5
10.5	10.5	27.0	27.0	35.0	35.0	57.5	57.5	72.0	72.0	72.0	72.0

72: C012006	73: C022006	74: C032006	75: C042006	76: C052006	77: C062006	78: C072006	79: C082006	80: C092006	81: C102006	82: C112006	83: C122006
Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
50.0	50.0	51.0	51.0	44.5	44.5	46.0	46.0	65.0	65.0	51.5	51.5
21.5	21.5	21.5	21.5	28.0	28.0	35.0	35.0	38.0	38.0	20.5	20.5
50.0	50.0	46.5	46.5	32.0	32.0	29.5	29.5	71.0	71.0	65.5	65.5
57.0	57.0	48.0	48.0	42.0	42.0	36.5	36.5	236.0	236.0	23.5	23.5

96: C012008	97: C022008	98: C032008	99: C042008	100: C052008	101: C062008	102: C072008	103: C082008	104: C092008	105: C102008	106: C112008	107: C122008
Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
54.0	54.0	46.5	46.5	49.5	49.5	65.5	65.5	76.5	76.5	48.5	48.5
42.5	42.5	23.5	23.5	19.0	19.0	26.5	26.5	29.5	29.5	12.5	12.5
31.0	31.0	42.0	42.0	37.0	37.0	67.0	67.0	84.0	84.0	41.0	41.0
0.0	0.0	0.0	0.0	0.0	0.0	62.5	62.5	16.0	16.0	0.0	0.0

108: C012009	109: C022009	110: C032009	111: C042009	112: C052009	113: C062009	114: C072009	115: C082009	116: C092009	117: C102009	118: C112009	119: C122009
Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
63.5	63.5	49.0	49.0	86.0	86.0	75.0	75.0	89.5	89.5	58.0	58.0
23.0	23.0	11.0	11.0	43.0	43.0	22.5	22.5	27.0	27.0	26.5	26.5
64.0	64.0	20.0	20.0	72.0	72.0	63.5	63.5	60.5	60.5	27.0	27.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	56.5	56.5

120: C012010	121: C022010	122: C032010	123: C042010	124: C052010	125: C062010	126: C072010	127: C082010	128: C092010	129: C102010	130: C112010	131: C122010
Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
75.0	75.0	69.5	69.5	55.0	55.0	55.0	55.0	93.5	93.5	69.0	69.0
19.5	19.5	19.5	19.5	27.0	27.0	26.5	26.5	30.0	30.0	24.0	24.0
45.0	45.0	45.0	45.0	38.5	38.5	30.5	30.5	48.0	48.0	42.0	42.0
14.0	14.0	23.0	23.0	33.0	33.0	8.5	8.5	30.5	30.5	18.0	18.0

132: C012011	133: C022011	134: C032011	135: C042011	136: C052011	137: C062011	138: C072011	139: C082011	140: C092011	141: C102011	142: C112011	143: C122011
Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
69.0	172.0	172.0	110.0	110.0	110.0	110.0	0.0	99.0	90.0	87.0	83.0
24.0	24.0	24.0	24.0	51.0	28.0	28.0	28.0	28.0	31.0	23.0	28.0
42.0	42.0	0.0	34.0	34.0	34.0	34.0	34.0	42.0	35.0	36.0	36.0
18.0	96.0	96.0	49.0	49.0	49.0	49.0	0.0	44.0	40.0	39.0	37.0

144: C012012	145: C022012	146: PAID_VOUCHERS_COUNT_PCT	147: PAYMENT_COUNT_PCT	148: PERSONS_BUILDING_CNT	149: HAS_LIC_FILE	150: **FRAUD**
Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal
78.0	0.0	16.0	14.0	18.0	1.0	NO
28.0	28.0	7.0	6.0	21.0	0.0	NO
36.0	37.0	100.0	39.0	21.0	0.0	NO
35.0	0.0	100.0	58.0	7.0	1.0	NO

Figura 4.21. Datos de entrada para la Detección de Fraude de 4 clientes

- Resultado

  El resultado que proporciona el Sistema es un listado de clientes con el idCliente, además la opción de ingresar el id del cliente para filtrar y realizar la búsqueda del cliente. La siguiente figura 4.22 muestra el resultado de la prueba de funcionalidad del sistema:

Figura 4.22. Resultado de la prueba de funcionalidad del sistema

4.5. Requerimientos para el desarrollo del Sistema

A continuación, se indican los requerimientos tanto a nivel de Hardware y Software, para el correcto desarrollo del Sistema Inteligente de detección de fraude de facturación de agua.

4.5.1. Requisitos mínimos a nivel de Hardware

- ✓ Capacidad de memoria Ram: 8 GB
- ✓ Procesador: Core i7de quinta generación 3.2. GHZ

- ✓ Capacidad Disco Duro: 500 GB

4.5.2. Requisitos mínimos a nivel de Software

Para la implementación del sistema, se escogió utilizar como lenguaje de programación en el lado del backend Python, que es un lenguaje de fácil uso, de un aprendizaje totalmente intuitivo y a la vez multiplataforma, el cual se puede usar tanto en Windows como en Linux, y sobre todo es Open Source.

- La tecnología Python ofrece muchas ventajas para un desarrollador de software, sobre todo usándolo conjuntamente con el framework Django:
  - Es una tecnología multiplataforma.
  - Framework MVC
  - Programación orientada a Objetos
- Para la interfaz del usuario, se usó la tecnología AngularJS, la cual contiene diversos componentes para las interfaces de usuario.
- Servidor propio para Django
- Para el desarrollo del modelo y su respectiva validación, se usó la librería weka 3.8 y para la aplicación del algoritmo de Árbol de decisión, Regresión Logísitca, LIBSVM y Vote, se instaló previamente las librerías correspondientes para cada uno de ellos.

CAPÍTULO 5: VALIDACIÓN DEL MODELO DE DETECCIÓN DE FRAUDE EN LA FACTURACIÓN DE AGUA

En el presente capítulo se muestra la validación del modelo híbrido propuesto, a travez de un dataset de la Municipalidad de Gaza, Palestina y del uso de los programas J48 (árbol de decisión), Logistic (Regresión Logística), LibSVM (Máquina de Soporte Vectorial) y el programa Vote (Modelo híbrido) de la herramienta Weka.

5.1. Diseño de validación

La validación sigue 4 pasos. Primero, se entrena el modelo propuesto con los datos de entrenamiento. Segundo, con el modelo entrenado se identifica los clientes fraudulentos respecto a los datos de validación. Si los resultados del paso 2 no son satisfactorios entonces se procede a calibrar los parámetros del modelo (Paso 3) y se retorna al paso 1, de lo contrario (satisactorios) el modelo propuesto es usado para identificar los clientes fraudulentos respecto a los datos de prueba (Paso 4). La calibración de los parámetros de los modelos que conforman el modelo propuesto se explica en la sección de entrenamiento (5.4). Ver figura 5.1.

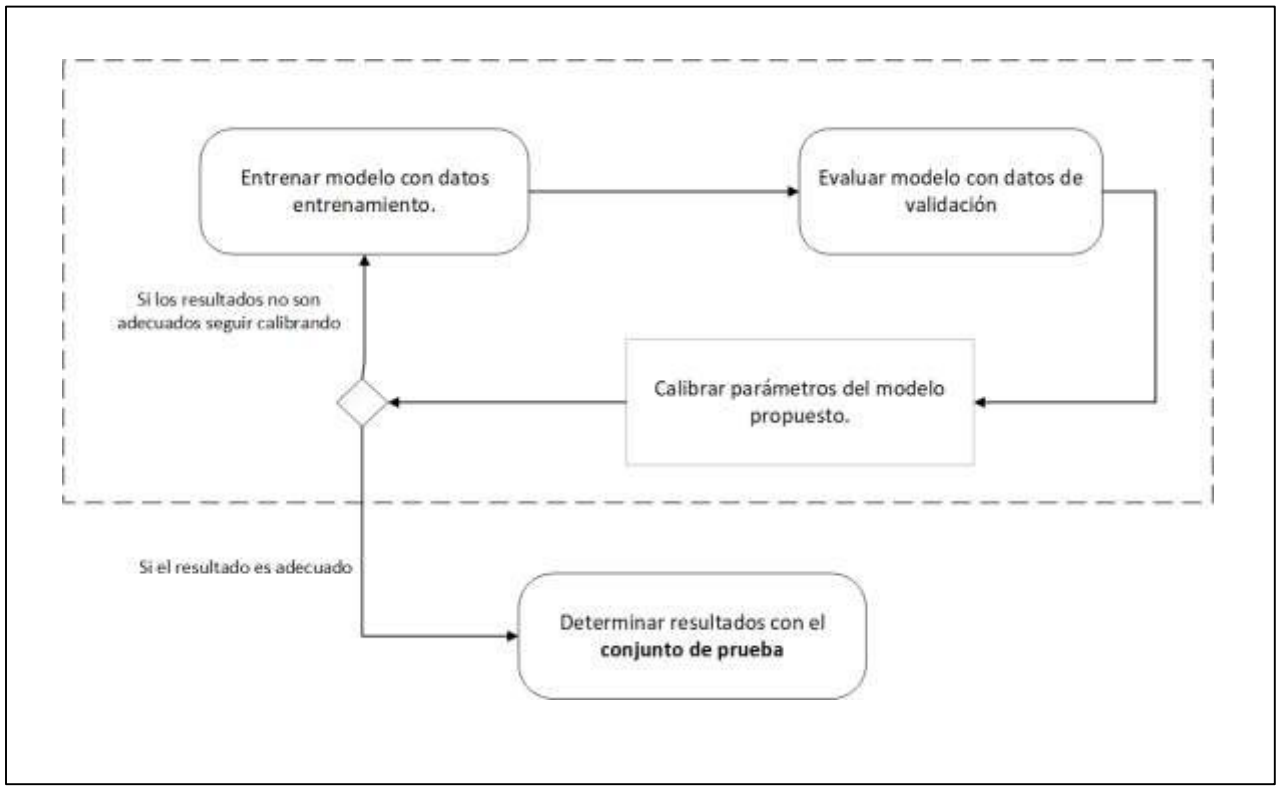


Figura 5.1. Diseño de la validación del modelo propuesto

Se ha seguido el diseño de evaluación dada en la figura 5.1 porque generó mejores resultados que el uso de la técnica cross validation (validación cruzada), los resultados de la validación cruzada se muetran en el ANEXO C.

Para la validaicón del modelo se usó los programas de la herrramienta Weka, J48, Logistic, LibSVM, Vote que implementan los clasificadores Árbol de Decisión, Regresión Logística, Máquina de Soporte Vectorial e Híbrido (Criterio de Votación) respectivamente.

5.2. Requerimientos para la operación del Sistema

5.2.1. Requerimientos mínimos a nivel de Hardware

- ✓ Capacidad de memoria Ram: 8 GB
- ✓ Procesador: Core i7de quinta generación 3.2. GHZ
- ✓ Capacidad Disco Duro: 500 GB

5.2.2. Requerimientos mínimos a nivel de Software

- ✓ Software Weka 3.8, package weka for Python (PYCHARM - JETBRAINS)
- ✓ Librería LibSVM, J48, Logistic y Vote
- ✓ Sistema Operativo: Windows 10

5.3. Instancias de pruebas

Se tiene un total de 4814 clientes, donde 4190 son clientes normales y el resto fraudulentos. Para ver todos los atributos y la dataset completa de prueba se puede acceder al siguiente link: https://bit.ly/2YCm06L

5.3.1. Preparación de datos

Se consdiero para los datos de entrenamiento, validación y pruebas las proporciones dadas en la tabla 5.1

	Entrenamiento	**Validación**	**Pruebas**	**Total**
Porcentaje	80%	10%	10%	100%
Registros	3852	481	481	4814

Tabla 5.1. Tabla de resultados según diferentes porcentajes de entrenamiento

Cada registro corresponde a un cliente y está conformado por 150 atributos (Ver ANEXO B).

5.3.2 Instancias de pruebas para el entrenamiento, validación y pruebas

La tabla 5.2, muestra la cantidad de instancias (registros) para el entrenamiento, validación y prueba, considerando registros normales y registros con fraude.

Instancias	Registros Totales	Registros Normales	Registros Fraude	% Registro Fraude
Entrenamiento	3852	3347	505	13,11%
Validación	481	405	76	15,80%
Pruebas	481	401	80	16,63%
Total	4814	4153	661	13,73%

Tabla 5.2. Instancias para el entrenamiento, validación y pruebas.

Los datos para el entrenamiento, validación y pruebas siguen la proporción dada en diversos trabajos (Usama, 1996):

- Para el entrenamiento, se considera el 80 %, que es un total de 3852 registros.
- Para la validación, se considera el 10 %, que es un total de 481 registros.
- Para las pruebas, se considera el 10 %, que es un total de 481 registros.

5.4. Métricas

Para realizar el análisis de los resultados del sistema de detección de fraude, se considera la matriz de consfusión, que según (Bense, 2007), es el instrumento más usual que se usa para evaluar la exactitud de una clasificación. Se construye la matriz de confusión como se observa en la tabla 5.3. En donde a, b, c y d representan respectivamente los números de "Clientes clasificados correctamente como fraudulentos", "Clientes fraudulentos clasificados incorrectamente como normales", "Clientes normales clasificados incorrectamente como fraudulentos" y "Clientes clasificados correcatamnete como normales".

Matriz de Confusión		Predicción	
		Cliente Fraude	Cliente Normal
REAL	Cliente Fraude	a	b
	Cliente Normal	c	d

Tabla 5.3. Matriz de confusión para el análisis de resultados.

Y se considera las siguientes métricas:

Sensibilidad (S): Nos indica de los clientes clasificados correctamente como fraudulentos respecto de los casos de clientes que cometen fraude, y se determina como:

$$\boldsymbol{S} = \frac{a}{a+b}$$

Especificad (E): Nos indica el número de casos clasificados correctamente como clientes normales respecto a cuantos se han identificado como normales por el sistema.

$$\boldsymbol{E} = \frac{d}{c+d}$$

Con lo cual se analiza la exactitud (Accuracy) de la siguiente manera:

- **Exactitud o Accuracy (AC):** Esta métrica nos indica el número total de clientes clasificados correctamente respecto al total de clientes:

$$\boldsymbol{AC} = \frac{a+d}{a+b+c+d}$$

5.5. Fase de entrenamiento

5.5.1. Calibración de parámetros

Para obtener un mejor resultado de entrenamiento del programa J48 se ha calibrado los parámetros ConfindenceFactor y MimNumObj, como sigue. ConfindenceFactor asume valores de 0.05 a 0.5 incrementando 0.05 por vez. MimNumObj asume valor de 1 ó 2. En la tabla 5.4 se muestran los resultados que se obtuvieron al variar los valores de cada uno de los parámetros para el programa J48.

MimNumObj: Mínimo número de instancias por hoja.

ConfindenceFactor: Umbral de confianza para la poda.

confindenceFactor	MimNumObj (Mínimo número de instancias por hoja)	Exactitud (%)
0.05	1	90.446
0.05	2	90.758
0.10	**1**	**91.227**
0.10	2	90.965
0.15	1	90.861
0.15	2	90.861
0.20	1	90.861
0.20	2	90.861
0.25	1	90.654
0.25	2	90.554
0.30	1	90.654
0.30	2	90.654
0.35	1	90.654
0.35	2	90.654
0.40	1	90.654
0.40	2	90.654
0.45	1	90.342
0.45	2	90.654
0.50	1	90.342
0.50	2	90.342

Tabla 5.4. Comparación de la exactitud del modelo variando el confidenceFactor y el MinNumObj del clasificador J48.

De la Tabla 5.4 podemos concluir que nuestro modelo obtiene una mejor tasa de exactitud alcanzando 91.227% cuando confidenceFactor es 0,10 y MimNumObj es 1.

Para obtener un mejor resultado de entrenamiento del programa Logistic se ha calibrado el parámetro maxlts, como sigue. Maxlts asume valores de -1 a 90. En la tabla 5.5 se muestran los resultados que se obtuvieron al variar el valor del parámetro para el programa Logistic.

maxllts: Número de iteracciones

maxlts (Número de iteracciones)	**Exactitud (%)**
-1	95.119
-2	86.396
0	86.396
1	88.265
2	91.277
3	92.253
4	93.042
5	93.769
10	93.769
15	94.081
20	94.600
35	95.119
40	94.807
45	94.117
50	95.015
60	95.223
70	**95.223**
80	95.119
90	95.119

Tabla 5.5. Comparación de la exactitud del modelo variando el maxlts del clasificador Logistic.

De la Tabla 5.5 podemos concluir que nuestro modelo obtiene una mejor tasa de exactitud alcanzando 97.223% cuando maxlts es 70.

Para obtener un mejor resultado de entrenamiento del programa LibSVM se ha calibrado los parámetros C (Costo) y Gamma, como sigue. C asume valores de 1 a 90 incrementando 10 por vez. Gamma varía su valor de 0.5 a 1. En la tabla 5.6 se muestran los resultados que se obtuvieron al variar los valores de sus parámetros para el programa LibSVM.

C: Complejidad del algortimo.

Gamma: Establece el valor de Gamma en la función del Kernel.

C(costo)	Gamma	Exactitud (%)
1	0.5	91.173
1	1	91.794
10	0.5	94.393
10	1	94.704
20	0.5	94.807
20	1	95.056
30	0.5	94.117
30	1	95.015
40	0.5	95.119
40	1	94.911
50	**0.5**	**95.327**
50	1	94.911
60	0.5	95.223
60	1	94.704
70	0.5	95.015
70	1	94.600
80	0.5	95.119
80	1	93.977
90	0.5	95.119
90	1	93.873
100	0.5	94.911
100	1	93.977

Tabla 5.6. Comparación de la exactitud del modelo variando el parámetro C y Gamma del algoritmo SVM.

De la Tabla 5.6 se concluye que nuestro modelo para LibSVM obtiene una mejor tasa de exactitud alcanzando 95.327% cuando C tiene un valor de 50 y gamma 0.5.

5.5.2. Configuración de Weka para el modelo híbrido

Para realizar el entrenamiento del modelo híbrido y obtener los resultados, se usó el 80 % de los registros y el resto se usó para la validación y pruebas. Este proceso se configurará con la herramienta Weka, tal como se ve en la figura 5.2.

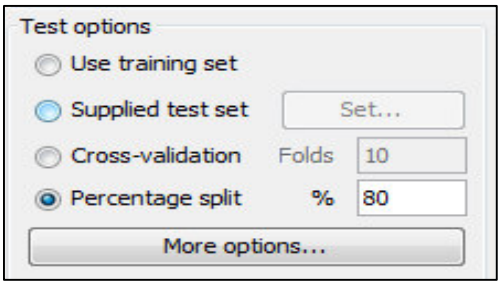


Figura 5.2. Configuración de porcentaje para el entrenamiento.

Se utilizará el clasificador 'Vote', que combina salidas de predicción de más de dos clasificadores. Se combinarán los clasificadores: Árbol de decisión (J48), Regresión logísitca (Logistic) y Máquina de soporte Vectorial (LibSVM), como se muestra en la figura 5.3 y 5.4:

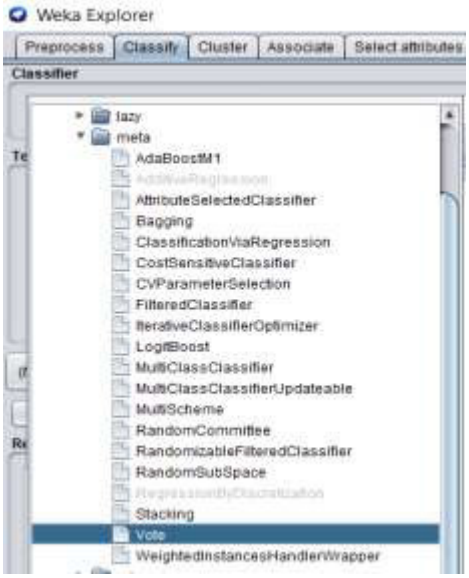


Figura 5.3. Selección del clasificador Vote

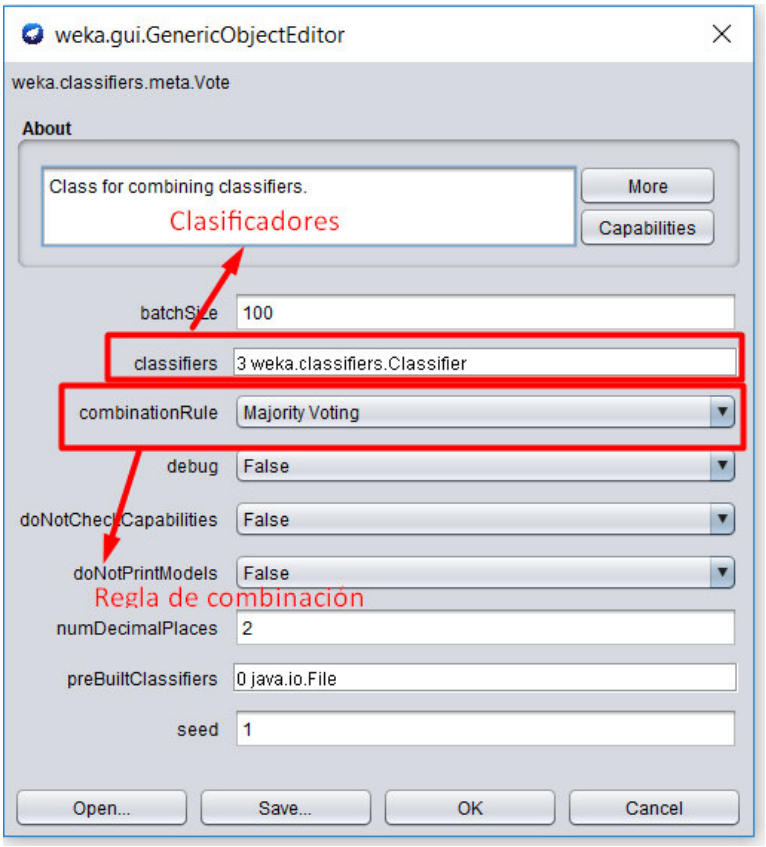


Figura 5.4. Configuración para ingresar los parámetros de 'Vote' en Weka

Como se ve en la figura 5.4, para la regla de combinación (combinationRule), se usará Voto Mayoritario (Majority Voting), y los clasificadores, las cuales se configurarán y se muestra respectivamente en la siguiente figura 5.5:

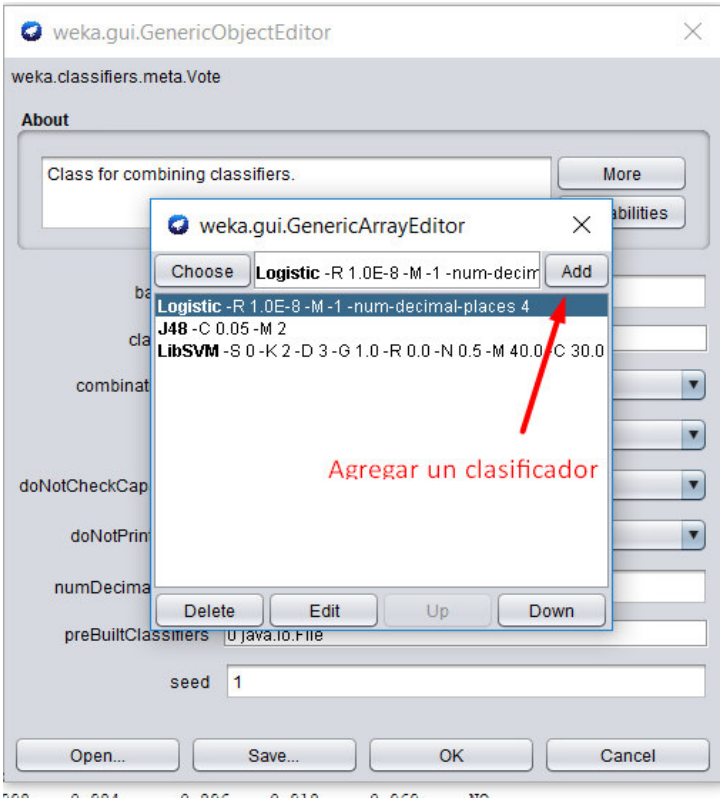


Figura 5.5. Configuración de clasificadores

Luego de haber realizado la configuración de los parámetros del clasificador Vote, seleccionamos la clase que se quiere predecir, en este caso FRAUD, damos clic en START, para empezar a entrenar nuestro modelo híbrido como se muestra en la siguiente figura 5.6:

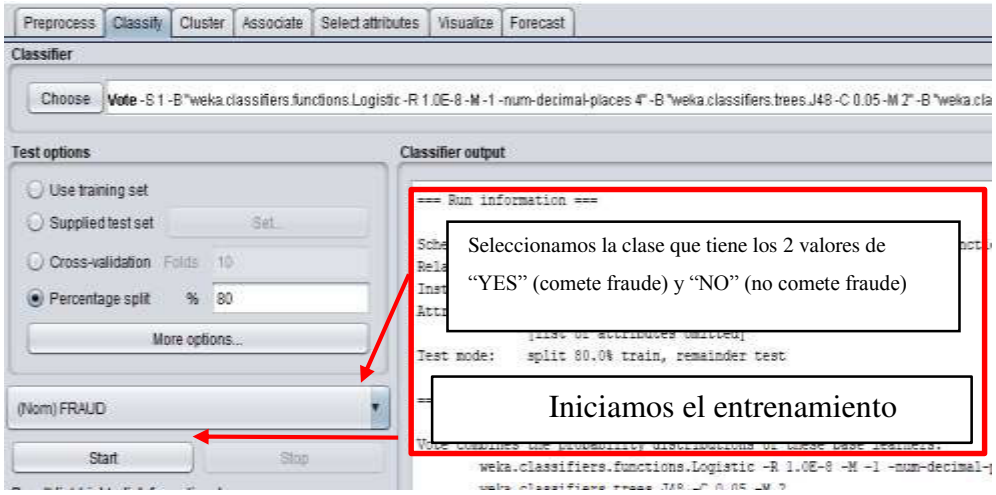
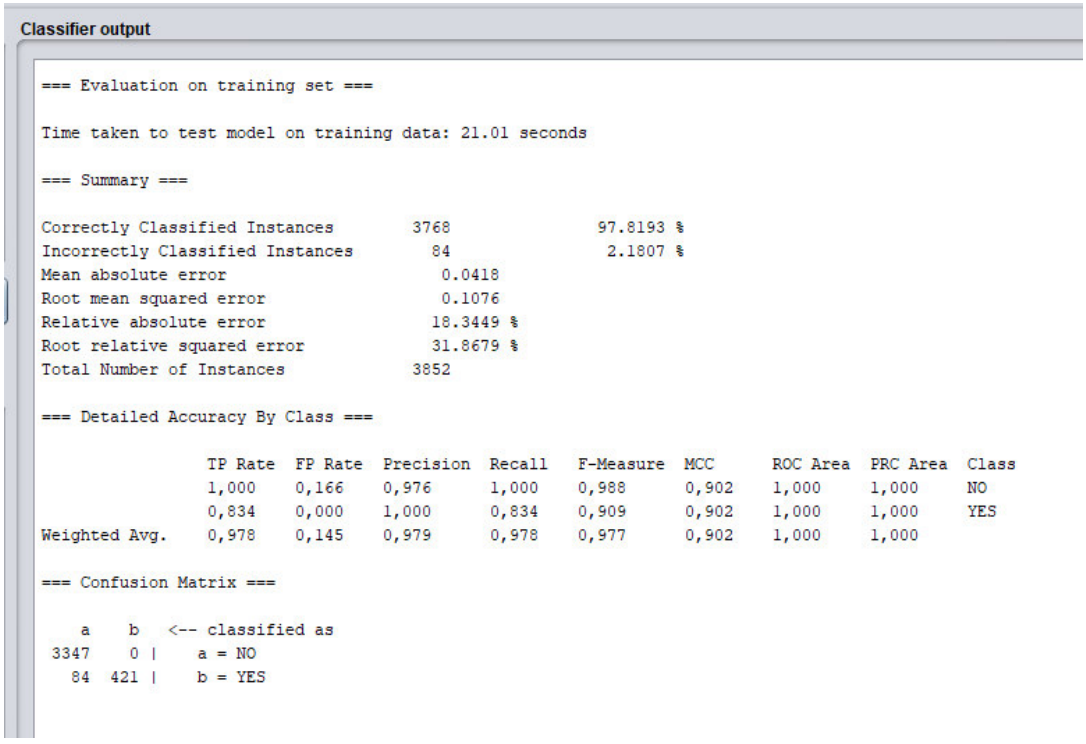


Figura 5.6. Iniciar el entrenamiento del modelo

5.5.3. Resultados

Luego de haber generado el modelo, obtenemos la siguiente salida con los respectivos resultados que se muestran en la figura 5.7:



Classifier output

```
=== Evaluation on training set ===

Time taken to test model on training data: 21.01 seconds

=== Summary ===

Correctly Classified Instances        3768               97.8193 %
Incorrectly Classified Instances        84                2.1807 %
Mean absolute error                      0.0418
Root mean squared error                  0.1076
Relative absolute error                 18.3449 %
Root relative squared error             31.8679 %
Total Number of Instances             3852

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1,000    0,166    0,976      1,000    0,988      0,902    1,000     1,000     NO
                 0,834    0,000    1,000      0,834    0,909      0,902    1,000     1,000     YES
Weighted Avg.    0,978    0,145    0,979      0,978    0,977      0,902    1,000     1,000

=== Confusion Matrix ===

    a    b   <-- classified as
 3347    0 |    a = NO
   84  421 |    b = YES
```

Figura 5.7. Resultados del entrenamiento

En la tabla 5.7 representamos la matriz de confusión del modelo híbrido entrenado:

Matriz de confusión		**PREDICCIÓN**	
		Cliente Fraude	Cliente Normal
REAL	Cliente Fraude	3347	0
	Cliente Normal	84	421

Tabla 5.7. Matriz de confusión del entrenamiento del modelo híbrido

Con lo cual se calcula la métrica de exactitud:

$$\textbf{Exactitud} = \frac{3347 + 421}{3347 + 0 + 84 + 421} = 0.978$$

Lo que nos arroja una tasa en la detección de un 97.81 %, y de los resultados de la detección obtenidas en el proceso de calibración (ver sección 5.5.1) se obtiene la siguiente tabla:

Clasificadores	**Exactitud (%)**
J48	91,22
Logistic	95,22
LibSVM	95,32
Vote	97,81

Tabla 5.8. Exactitud de modelos y el clasificador Vote.

5.6. Fase de validación

Para validar el modelo entrenado, debemos seguir los siguientes pasos:

Paso 1: Cargar nuestro modelo entrenado, tal como se muestran en las siguientes figuras 5.8 y 5.9:

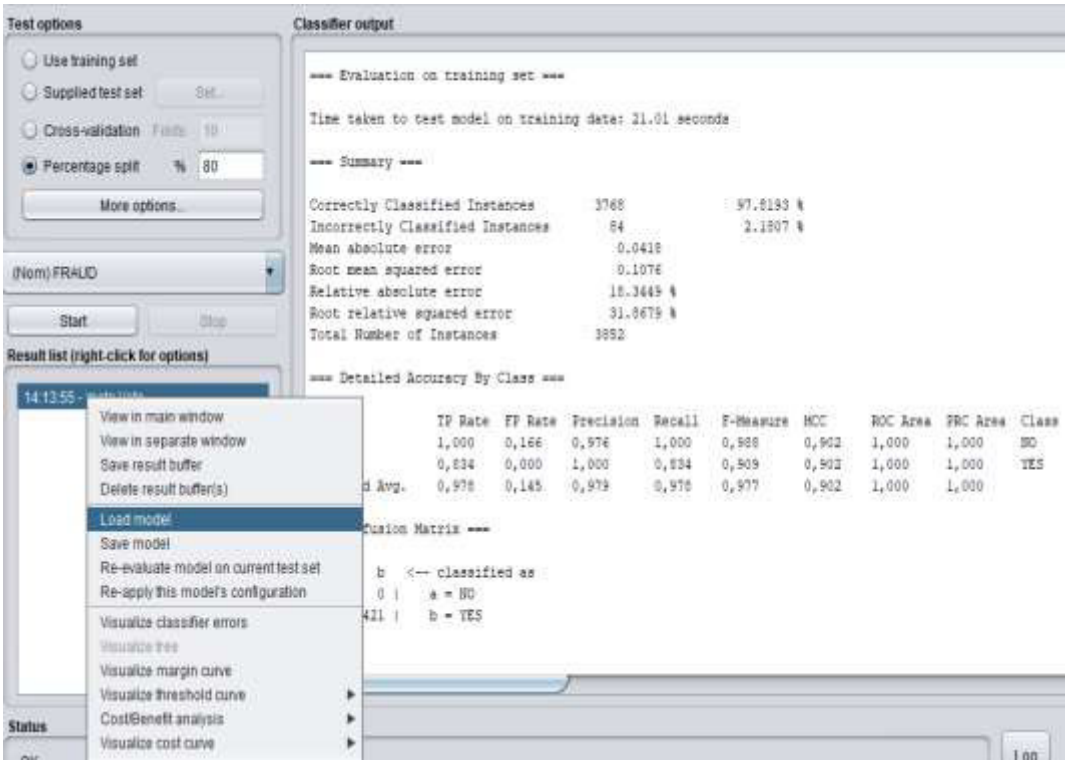


Figura 5.8. Cargar Modelo entrenado.

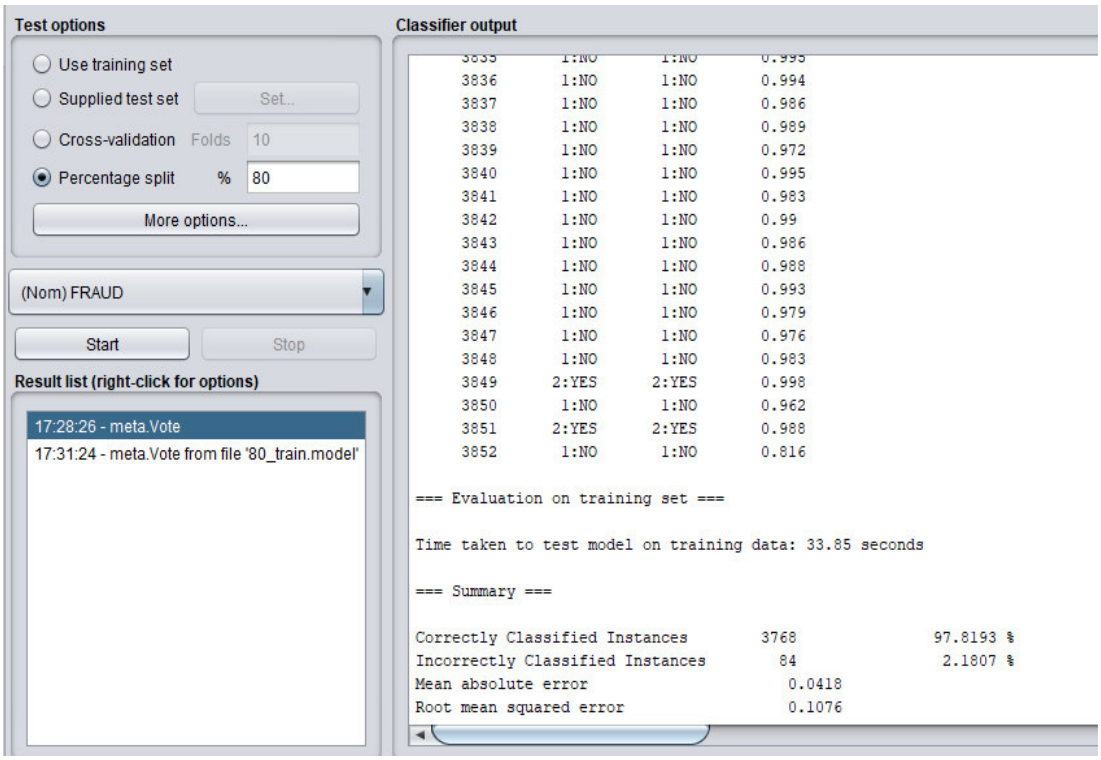


Figura 5.9. Modelo entrenado ha sido cargado

Paso 2: Seguidamente, cargamos el conjunto de instancias para la validación que son un total de 481 registros, tal como se muestra en la figura 5.10:

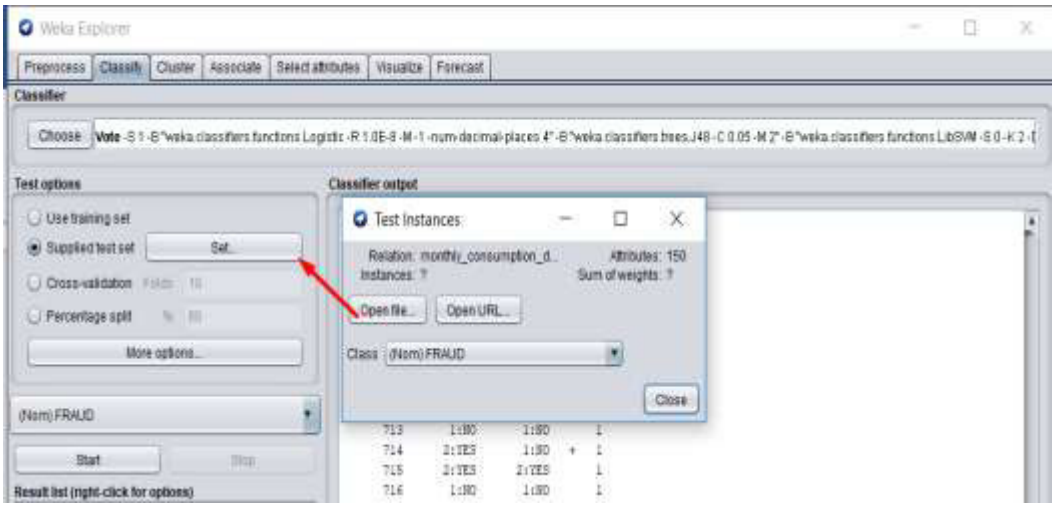


Figura 5.10. Carga de datos para la validación

Paso 3: Validar sobre este conjunto de datos. Para esto damos clic derecho sobre el modelo que fue cargado y seleccionamos la opción ‘Re-evalute on current test set’, tal como se muestra en la figura 5.11:

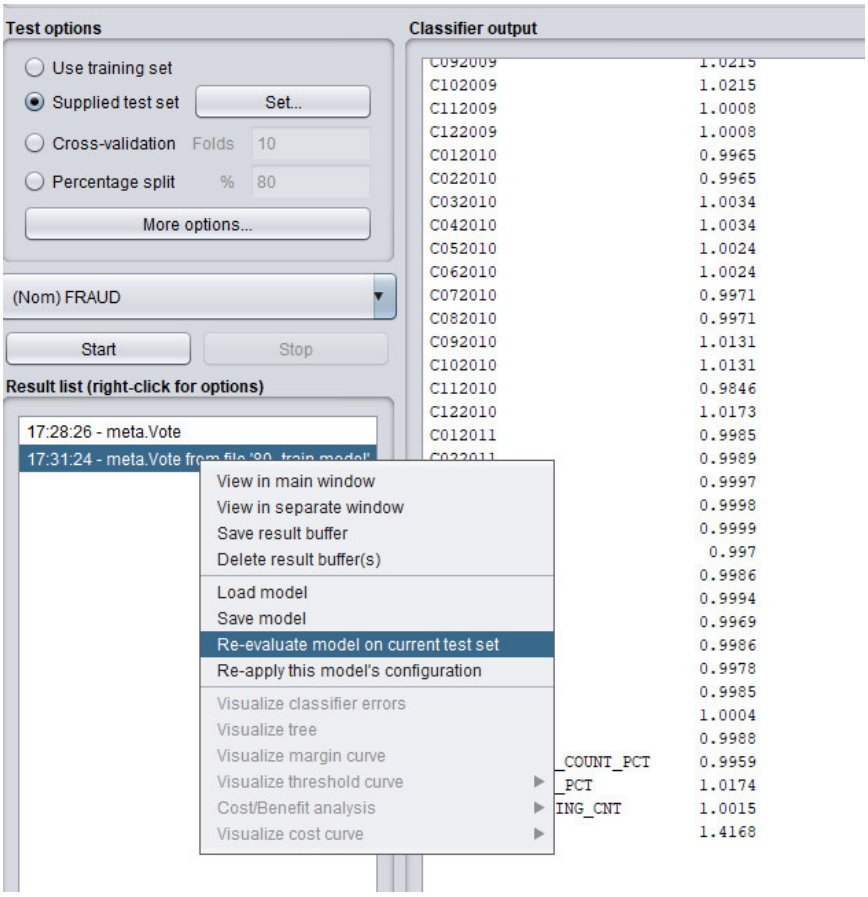


Figura 5.11. Validación de los datos sobre el modelo creado

Finalmente, se obtendrá los siguientes resultados que se muestran en la figura 5.12:

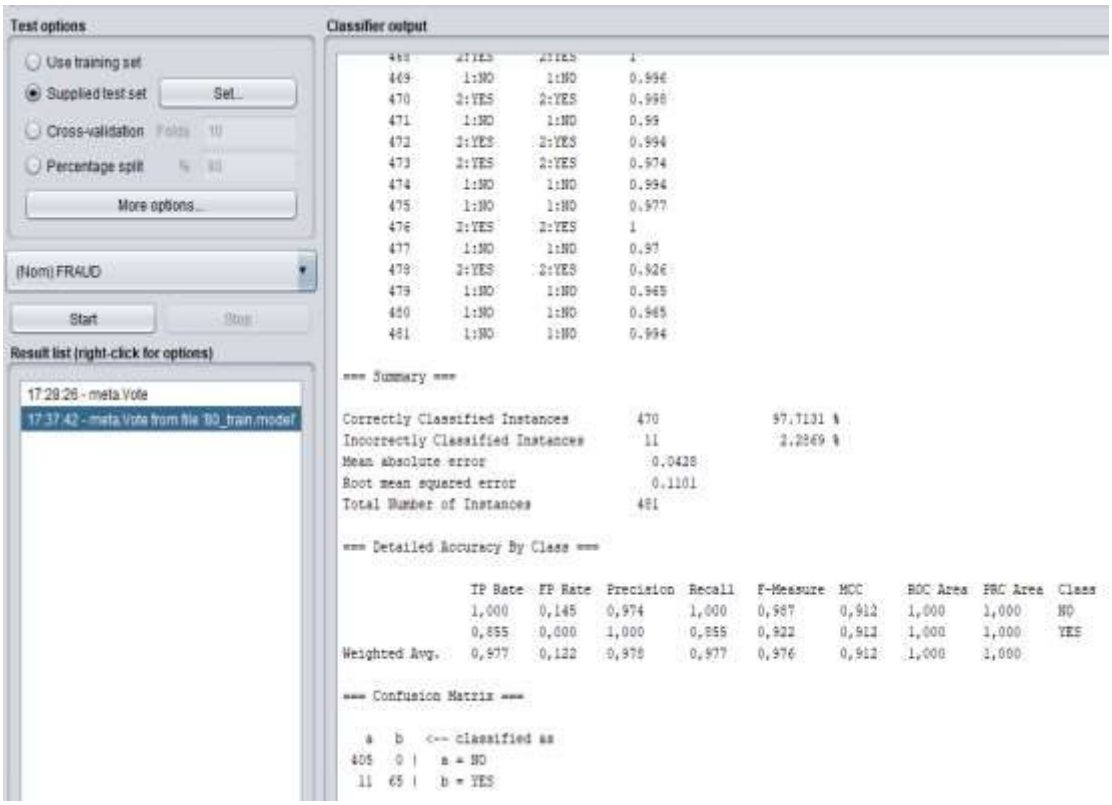


Figura 5.12. Resultados de la validación del Modelo Híbrido

5.7. Análisis de los resultados de la validación del modelo

Para realizar el análisis de la tasa de exactitud del sistema de detección de fraude, se construye la matriz de confusión a partir de los resultados generados por Weka tal como se observa en la tabla 5.9, la cual nos indica la cantidad de clientes detectados como fraudulentos de manera correcta o no, para su respectiva validación.

Matriz de Confusión		Predicción	
		Cliente Fraude	Cliente Normal
REAL	Cliente Fraude	405	0
	Cliente Normal	11	65

Tabla 5.9. Matriz de confusión, detección del sistema vs. Detección real

Seguidamente se analizará los resultados para cada uno de las métricas:

- **Sensibilidad (S):** Hallando la sensibilidad tenemos:

$$\boldsymbol{S} = \frac{405}{405 + 0}$$

El resultado es "1" que es equivalente a 100 %

- **Especificad (E):** Hallando la especificidad nos resulta:

$$\boldsymbol{E} = \frac{65}{11 + 65}$$

El resultado es "0.855" que es equivalente a 85,5 %

- **Exactitud o Accuracy (AC):**

$$\boldsymbol{AC} = \frac{405 + 65}{405 + 0 + 11 + 65}$$

La cual nos arroja una tasa de un 97.71 %

5.8. Confirmación de los resultados.

Para confirmar los resultados de la validación se procederá a usar el conjunto de pruebas que representa el 10% del total de registros, que son 481 instancias. Para ello procedemos a ejecutar los siguientes pasos:

Paso 1: Cargar nuestro conjunto de pruebas, tal como se muestran en la siguiente figura 5.13

Figura 5.13. Cargamos nuestro conjunto de pruebas

Paso 2: Probamos sobre este conjunto de datos. Para esto damos clic derecho sobre el modelo que fue cargado y seleccionamos la opción ‘Re-evalute on current test set’, tal como se muestra en la figura 5.14:

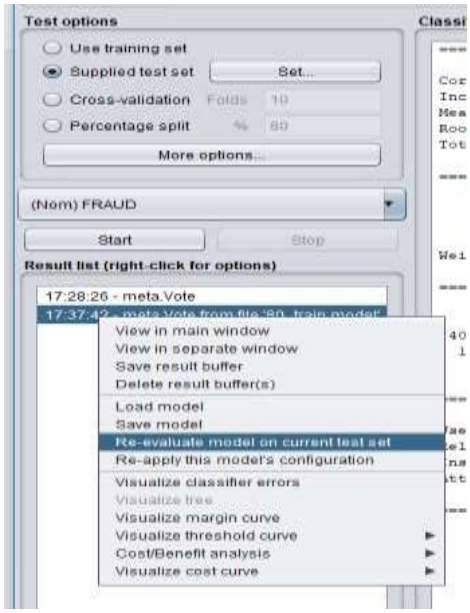
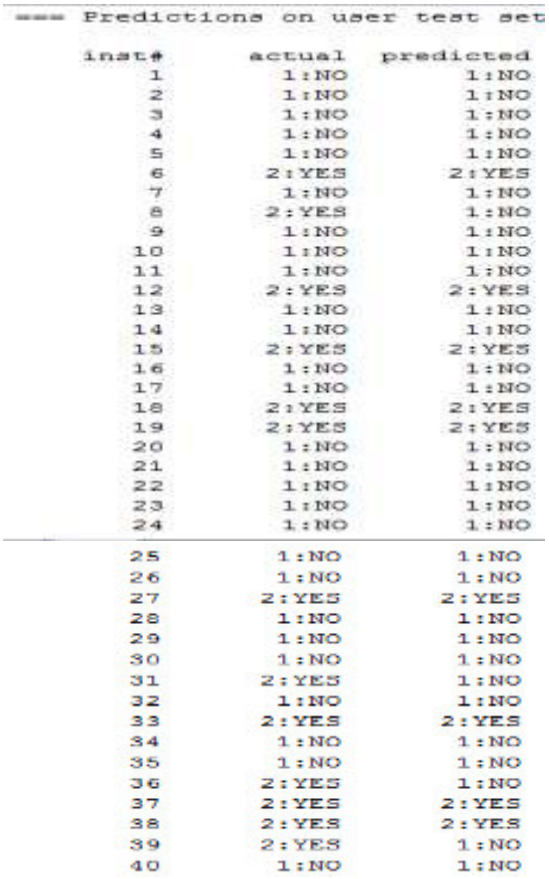


Figura 5.14. Confirmación sobre los resultados.

Finalmente, se obtendrá los resultados para 481 instancias el cual se presenta en el Anexo D, los primeros 40 registros de este anexo se muestra en la figura 5.15.



=== Predictions on user test set

inst#	actual	predicted
1	1:NO	1:NO
2	1:NO	1:NO
3	1:NO	1:NO
4	1:NO	1:NO
5	1:NO	1:NO
6	2:YES	2:YES
7	1:NO	1:NO
8	2:YES	1:NO
9	1:NO	1:NO
10	1:NO	1:NO
11	1:NO	1:NO
12	2:YES	2:YES
13	1:NO	1:NO
14	1:NO	1:NO
15	2:YES	2:YES
16	1:NO	1:NO
17	1:NO	1:NO
18	2:YES	2:YES
19	2:YES	2:YES
20	1:NO	1:NO
21	1:NO	1:NO
22	1:NO	1:NO
23	1:NO	1:NO
24	1:NO	1:NO
25	1:NO	1:NO
26	1:NO	1:NO
27	2:YES	2:YES
28	1:NO	1:NO
29	1:NO	1:NO
30	1:NO	1:NO
31	2:YES	1:NO
32	1:NO	1:NO
33	2:YES	2:YES
34	1:NO	1:NO
35	1:NO	1:NO
36	2:YES	1:NO
37	2:YES	2:YES
38	2:YES	2:YES
39	2:YES	1:NO
40	1:NO	1:NO

Figura 5.15. Resultado para las primeras 40 instancias del conjunto de pruebas.

Error = 7.48%

Exactitud = 92.52%

CAPÍTULO 6: CONCLUSIONES Y TRABAJOS FUTUROS

6.1. Conclusiones

6.1.1. Conclusión general

Se implementó un sistema inteligente usando técnicas de minería de datos: Árbol de decisión, Regresión Logística y Máquina de Soporte Vectorial, que mediante características y comportamiento de un cliente de consumo de agua se logra detectar si éste comete fraude en su facturación mensual o no.

6.1.2. Conclusiones específicas

6.1.2.1. Objetivo específico 1

Se realizó una revisión de la literatura y se encontró 1 modelo, 6 técnicas de Machine Learning, Árbol de Decisión, Regresión Logística, Redes Neuronales, K-Nearest, Detección de datos atípicos, Coeficiente de Person, Algoritmos estadísticos, para detección de fraude en consumo de agua en la revisión de la literatura en el capítulo 2, y se llegó a la conclusión que existe una mejor exactitud en la detección de fraude al combinar 3 a más técnicas.

6.1.2.2. Objetivo específico 2

Se diseñó un modelo híbrido de minería de datos basado en el clasificador 'Vote', que mezcla los resultados de los clasificadores: Árbol de decisión, Regresión logística y Máquina de Soporte Vectorial; que considera las variables que se identificaron en la literatura y que se encontraron en diversos estudios, para la construcción del modelo y que se visualiza en la figura 3.1.

6.1.2.3. Objetivo específico 3

Se desarrolló un sistema web basado en el lenguaje de programación Python, el cual se realizó con el software Weka y la librería 'Vote' para las tareas de entrenamiento y validación del modelo híbrido final incluyendo la detección de clientes que cometen fraude en su facturación de agua, tal como se aprecia en la figura 4.22, lo cual lo convierte en un sistema inteligente.

6.1.2.4. Objetivo específico 4

Se ha realizado el entrenamiento y validación del modelo de detección de fraude usando la data histórica de consumo de clientes normales y fraudulentos de un dataset de la Municipalidad de Gaza, en Palestina. La tasa de exactitud del modelo obtenido luego de la validación es de 97.713 %, tal como se muestra en la figura 5.12, lo cual es una mejor tasa que se alcanzó en el estudio realizado por (Palomino y Rivera, 2016) donde obtuvieron una tasa de 95.7 %. El cual favorece también en el proceso de la toma de decisiones de identificación de fraude y económicamente en las utilidades que percibe la empresa y sus ejecutivos, al empezar a cobrar el agua no facturada por estas personas que cometen el fraude y que prevalecía por mucho tiempo sin resolver.

6.2. Limitaciones

- El trabajo de investigación solo se orienta para personas naturales, más no a empresas que consumen agua potable.
- Para que este sistema de detección de fraudes pueda entrenar y validar, se requiere de los datos históricos de consumo de agua potable mensual del cliente.

6.3. Trabajos futuros

- Según los antecedentes expuestos, existen varias empresas a nivel nacional e internacional prestadoras de servicios de Agua Potable que tienen el mismo problema presentado, por lo tanto, es posible plantearles dicha solución a estas empresas.
- El sistema propuesto e implementado puede integrarse a servicios web, de los sistemas propietarios por las empresas que brindan servicios de agua potable y tengan la posibilidad de integrarlo a sus sistemas.
- La solución planteada también puede aplicarse en empresas de diferentes servicios como empresas que suministran energía eléctrica y tengan el mismo problema de fraude en su facturación de consumo.
- Según las pruebas hechas con otros modelos de entrenamiento, se pueden incluir más técnicas de Machine Learning, como Redes Bayesianas, Regresión Lineal, entre otras, y finalmente obtener un híbrido que llegue a un mínimo de 97 % de detección.

REFERENCIAS BIBLIOGRÁFICAS

Peleg, A., Armon, A., Barkay, U., Scolnicov, H., & Gutner, S. (2011). System and method for monitoring resources in a water utility network. *United States Patent*.

Patiño Espinoza, V. (2014). Modelo de detección de fraude en clientes del servicio de agua potable de una empresa sanitaria. Santiago de Chile, Chile: UNIVERSIDAD DE CHILE.

Diaz, R. E. (2008). Creación y aplicación de un programa de control de fraudes domiciliarios para minimizar las pérdidas de agua potable en aguas del altiplano s.a. (Tesis de Pregrado). Santiago de Chile, Chile.

Monedero, I., Biscarri, F., Guerrero, J. I., Roldán, M., & León, C. (2015). An Approach to Detection of Tampering in Water Meters. *Procedia Computer Science*, 413-421.

Humaid, E. H., & Barhoum, T. (2013). Water Consumption Financial Fraud Detection: A Model Based on Rule Induction. *Palestinian International Conference on Information and Communication Technology*, (págs. 115-120). Gaza.

Richard, P., & Mijail, R. (2016). Sistema Inteligente para detectar fraude en el servicio de Agua Potable de una Empresa Sanitaria. Caso de Estudio: Municipalidad de Gaza, Palestina. (Tesis de Pregrado). Universidad Nacional Mayor de San Marcos, Lima, Perú.

Candelieri, A. (2017). Clustering and Support Vector Regression for Water Demand Forecasting and Anomaly Detection. *Water*, (págs. 9-224).

Fettermann, D. d., Guerra, K. C., Mano, A. P., & Marodim, G. A. (2015). Uma sistemática para detecção de fraudes em empresas de abastecimento de água. *Interciencia*, 114-120.

Maghrebi, M., Aghaebrahimi, M. R., Taherian, H., & Attari, M. (2014). Determining the amount and location of leakage in water supply networks using a neural network improved by the Bat optimization algorithm. *Civil Engineering Urbanism*, 322-327.

Gagliardi, F., Alvisi, S., Kapelan, Z., & Franchini, M. (2017). A Probabilistic Short-Term Water Demand Forecasting Model Based on the Markov Chain. *Water*, 9-507.

Naib, M., & Chhabra, A. (2017). Ensemble vote approach for predicting primary tumors using data mining. *Confluence*, (págs. 97-102).

Leyva, Y. O., Vázquez, E. A., & Boada, D. H. (2011). Integración entre python y Weka aplicado en la minería de datos. *V Taller de Calidad en las Tecnologías de la Información y las Comunicaciones.*

Cabral, J. E., Pinto, J. O., Martins, E. M., & Pinto, A. M. (2008). Fraud detection in high voltage electricity consumers using data mining. *IEEE/PES Transmission and Distribution Conference and Exposition*, (págs. 1-5). Chicago.

Spirić, J. &., Dočić, S. &., & Miroslav & D. Popović, T. (2014). Using the rough set theory to detect fraud committed by electricity customers. *International Journal of Electrical Power & Energy Systems*, 727-734.

Ding, N., Gao, H., Bu, H., Ma, H., & Si, H. (2018). Multivariate-Time-Series-Driven Real-time Anomaly Detection Based on Bayesian Network. *Sensors*.

Moreno Palenzuela, J. (2018). Regresión logística basada en distancias para detección de fraude en el IRPF. (Tesis de Maestría). Madrid, España.

Candelieri, A., Soldi, D., & Archetti, F. (2015). Short-term forecasting of hourly water consumption by using automatic metering readers data. *Procedia Engineering*, 844-853.

Landa-Torres, F. &.-G., Rebolledo-Mendez, S. &., Brizio, G. &., & Héctor & Huerta - Pacheco, N. (2015). Evaluador de Eficiencias de Técnicas de Clasificación en R. *Foro Nacional de Estadística*, (págs. 66-70). At Aguascalientes.

Sedapal, Servicio de Agua Potable y Alcantarillado - (SEDAPAL S.A.). (19 de Febrero de 2019). Obtenido de http://www.aai.com.pe/wp-content/uploads/2018/10/Sedapal-Jun-18rev.pdf.

Errecalde, M. L. (2019). Preparación de los datos en el proceso KDD.

Coma-Puig, B., Carmona, J., Gavaldá, R., & Martin, S. A. (2016). Fraud Detection in Energy Consumption: A Supervised Approach. *EEE International Conference on Data Science and Advanced Analytics (DSAA)*, (págs. 120-129). Montreal.

Mounce, S. R., Mounce, R. B., Jackson, T., Austin, J., & Boxall, J. B. (2014). Pattern matching and associative artificial neural networks for water distribution system time series data analysis. *Journal of Hydroinformatics*, 617-632.

Al-Radaideh, Q. A., & Al-Zoubi, M. M. (2018). A data mining based model for detection of fraudulent behaviour in water consumption. *International Conference on Information and Communication Systems (ICICS)*, (págs. 48-54). Irbid.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence*, 37-53.

Pyle, D. (1999). *Data preparation for data mining.* San Francisco.

Willmott, J., & C & Matsuura, K. (2005). Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance. *Climate research*, 79-82.

Rey, D. P. (2007). Un modelo de integración y preprocesamiento de información. (Tesis Doctoral). Madrid.

Vigentes, A. A.-T. (27 de Diciembre de 2012). Obtenido de http://www.siss.gob.cl/586/articles-4625_aguas_andinas_g1_sep2012.pdf

LEDESMA, O. (2011). *Reducción de Pérdida por Micromedición.*

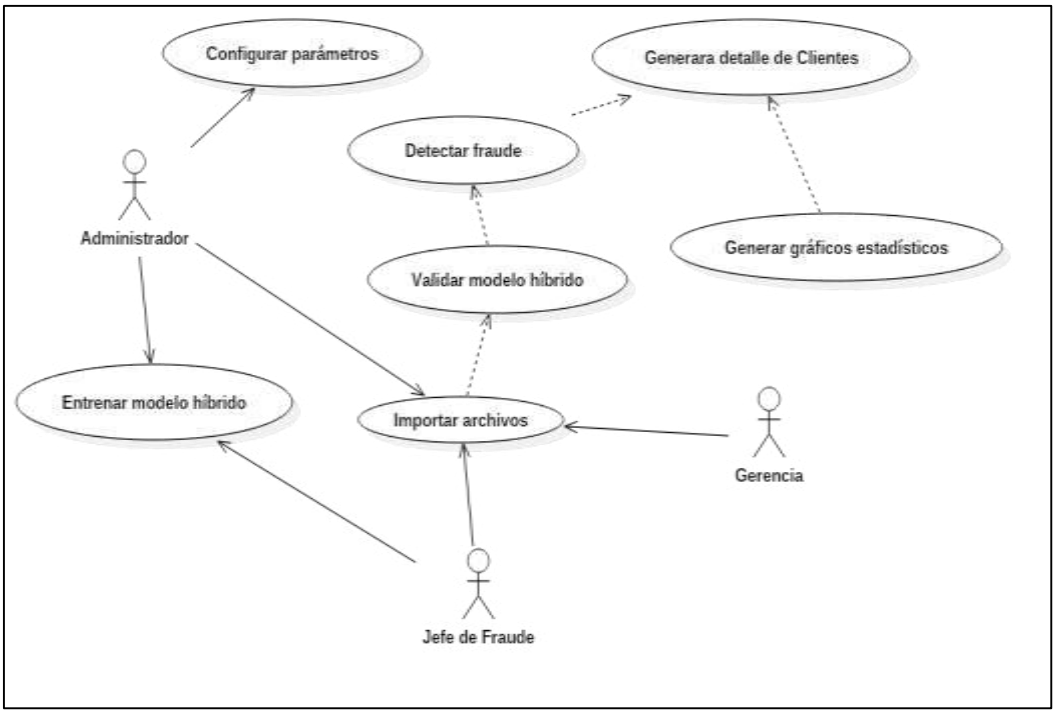
McKenzie, A. L. (2009). Ten years in using the UARL formula to calculate Infrastructure. *International Water Data Comparisons Ltd.* South Africa.

ANEXO A

A.1. Casos de uso del Sistema Inteligente

Paquete CUS	Nombre CUS	Descripción
Acceso al sistema	Validar Usuario	Este CUS permite validar el acceso de los usuarios al sistema, con la respectiva clave y contraseña asiganda a un usuario.
Mantenimiento del Modelo Híbrido	Importar Archivo	Este CUS permite al usuario importar un archivo de formato CSV ya sea para entrenamiento o validación del modelo.
	Entrenar Modelo	Este CUS permite generar el archivo de entrenamiento del modelo híbrido para luego ser validado.
	Validar Modelo	Este CUS realiza la validación del modelo híbrido entrenado previamente, confirmando la tasa de exactitud para la detección de fraude.
Detección de fraude	Detectar Fraude	Este CUS permite analizar los registros cargados y la visualización de clientes que están cometiendo fraude.
	Generar detalle de Clientes	Este CUS permite visualizar la información requerida de todos los clientes que han cometido fraude para su respectivo análisis.

A.2. Diagrama de Casos de Uso del Sistema



ANEXO B

B.1. Descripción de las variables usadas en la dataset

Variable	Descripción
AGREEMENT ID	Identificador del acuerdo, es el id con el cual se le puede identificar al cliente.
CXXYYYY	Es el consumo (m^3) del cliente en el mes XX del año YYYY.
PAID_VOUCHERS_COUNT_PCT	Es el número de facturas que han sido pagados totalmente por los clientes.
PAYMENT_COUNT_PCT	Es el porcentaje de la cantidad de pagos de los clientes con respecto a la cantidad de visitas para que realice el pago de las facturas
Fraud	Indica si el cliente esta categorizado como fraudulento o normal. Toma dos valores YES/NO.

En la tabla 7.2, la variable CXXYYY va desde C032000 (marzo del año 2000) hasta C022012 (febrero del año 2012).

ANEXO C

C.1. Resultados aplicando la técnica de Cross Validation (Validación cruzada)

Folds	Exactitud (%)
5	92.77
6	93.22
7	92.95
8	93.12
9	92.99
10	92.93
15	93.24
20	93.29

ANEXO D

D.1. Resultados para todos los registros en la confirmación de los resultados

inst#	actual	predicted
1	1:NO	1:NO
2	1:NO	1:NO
3	1:NO	1:NO
4	1:NO	1:NO
5	1:NO	1:NO
6	2:YES	2:YES
7	1:NO	1:NO
8	2:YES	1:NO
9	1:NO	1:NO
10	1:NO	1:NO
11	1:NO	1:NO
12	2:YES	2:YES
13	1:NO	1:NO
14	1:NO	1:NO
15	2:YES	2:YES
16	1:NO	1:NO
17	1:NO	1:NO
18	2:YES	2:YES
19	2:YES	2:YES
20	1:NO	1:NO
21	1:NO	1:NO
22	1:NO	1:NO
23	1:NO	1:NO
24	1:NO	1:NO
25	1:NO	1:NO
26	1:NO	1:NO
27	2:YES	2:YES
28	1:NO	1:NO
29	1:NO	1:NO
30	1:NO	1:NO
31	2:YES	1:NO
32	1:NO	1:NO
33	2:YES	2:YES
34	1:NO	1:NO
35	1:NO	1:NO
36	2:YES	1:NO
37	2:YES	2:YES
38	2:YES	2:YES
39	2:YES	1:NO
40	1:NO	1:NO
41	1:NO	1:NO
42	2:YES	1:NO
43	1:NO	1:NO
44	1:NO	1:NO
45	1:NO	1:NO
46	1:NO	1:NO
47	1:NO	1:NO
48	2:YES	1:NO
49	1:NO	1:NO
50	1:NO	1:NO

50	1:NO	1:NO
51	1:NO	1:NO
52	1:NO	1:NO
53	1:NO	1:NO
54	1:NO	1:NO
55	2:YES	1:NO
56	1:NO	1:NO
57	1:NO	1:NO
58	1:NO	1:NO
59	1:NO	1:NO
60	1:NO	1:NO
61	1:NO	1:NO
62	1:NO	1:NO
63	1:NO	1:NO
64	2:YES	2:YES
65	1:NO	1:NO
66	1:NO	1:NO
67	1:NO	1:NO
68	1:NO	1:NO
69	1:NO	1:NO
70	1:NO	1:NO
71	1:NO	1:NO
72	1:NO	1:NO
73	2:YES	1:NO
74	1:NO	1:NO
75	2:YES	2:YES
76	1:NO	1:NO
77	1:NO	1:NO
78	1:NO	1:NO
79	1:NO	1:NO
80	1:NO	1:NO
81	1:NO	1:NO
82	2:YES	2:YES
83	1:NO	1:NO
84	2:YES	1:NO
85	1:NO	1:NO
86	1:NO	1:NO
87	1:NO	1:NO
88	1:NO	1:NO
89	1:NO	1:NO
90	1:NO	1:NO
91	1:NO	1:NO
92	1:NO	1:NO
93	1:NO	1:NO
94	1:NO	1:NO
95	1:NO	1:NO
96	1:NO	1:NO
97	1:NO	1:NO
98	1:NO	1:NO
99	1:NO	1:NO
100	1:NO	1:NO

101	1:NO	1:NO
102	1:NO	1:NO
103	2:YES	2:YES
104	1:NO	1:NO
105	2:YES	2:YES
106	1:NO	1:NO
107	1:NO	1:NO
108	1:NO	1:NO
109	1:NO	1:NO
110	1:NO	1:NO
111	1:NO	1:NO
112	1:NO	1:NO
113	1:NO	1:NO
114	1:NO	1:NO
115	1:NO	1:NO
116	1:NO	1:NO
117	2:YES	1:NO
118	2:YES	2:YES
119	1:NO	1:NO
120	1:NO	1:NO
121	1:NO	1:NO
122	1:NO	1:NO
123	1:NO	1:NO
124	1:NO	1:NO
125	1:NO	1:NO
126	1:NO	1:NO
127	1:NO	1:NO
128	1:NO	1:NO
129	1:NO	1:NO
130	2:YES	2:YES
131	1:NO	1:NO
132	1:NO	1:NO
133	1:NO	1:NO
134	1:NO	1:NO
135	2:YES	1:NO
136	1:NO	1:NO
137	1:NO	1:NO
138	1:NO	1:NO
139	1:NO	1:NO
140	2:YES	1:NO
141	1:NO	1:NO
142	1:NO	1:NO
143	1:NO	1:NO
144	1:NO	1:NO
145	1:NO	1:NO
146	1:NO	1:NO
147	1:NO	1:NO
148	1:NO	1:NO
149	1:NO	1:NO
150	1:NO	1:NO

151	1:NO	1:NO
152	1:NO	1:NO
153	1:NO	1:NO
154	1:NO	1:NO
155	1:NO	1:NO
156	2:YES	2:YES
157	1:NO	1:NO
158	1:NO	1:NO
159	2:YES	2:YES
160	1:NO	1:NO
161	2:YES	1:NO
162	1:NO	1:NO
163	1:NO	1:NO
164	2:YES	2:YES
165	1:NO	1:NO
166	1:NO	1:NO
167	2:YES	1:NO
168	1:NO	1:NO
169	1:NO	1:NO
170	2:YES	2:YES
171	1:NO	1:NO
172	1:NO	1:NO
173	1:NO	1:NO
174	1:NO	1:NO
175	1:NO	1:NO
176	1:NO	1:NO
177	2:YES	2:YES
178	1:NO	1:NO
179	2:YES	2:YES
180	1:NO	1:NO
181	1:NO	1:NO
182	1:NO	1:NO
183	1:NO	1:NO
184	1:NO	1:NO
185	1:NO	1:NO
186	1:NO	1:NO
187	1:NO	1:NO
188	1:NO	1:NO
189	2:YES	1:NO
190	1:NO	1:NO
191	2:YES	1:NO
192	1:NO	1:NO
193	1:NO	1:NO
194	1:NO	1:NO
195	2:YES	2:YES
196	2:YES	2:YES
197	2:YES	2:YES
198	1:NO	1:NO
199	2:YES	2:YES
200	1:NO	1:NO

201	1:NO	1:NO
202	1:NO	1:NO
203	1:NO	1:NO
204	1:NO	1:NO
205	1:NO	1:NO
206	1:NO	1:NO
207	1:NO	1:NO
208	1:NO	1:NO
209	2:YES	1:NO
210	1:NO	1:NO
211	2:YES	2:YES
212	2:YES	1:NO
213	1:NO	1:NO
214	1:NO	1:NO
215	1:NO	1:NO
216	1:NO	1:NO
217	1:NO	1:NO
218	1:NO	1:NO
219	2:YES	1:NO
220	1:NO	1:NO
221	1:NO	1:NO
222	1:NO	1:NO
223	1:NO	1:NO
224	1:NO	1:NO
225	1:NO	1:NO
226	1:NO	1:NO
227	1:NO	1:NO
228	1:NO	1:NO
229	1:NO	1:NO
230	2:YES	1:NO
231	1:NO	1:NO
232	1:NO	1:NO
233	1:NO	1:NO
234	1:NO	1:NO
235	1:NO	1:NO
236	1:NO	1:NO
237	1:NO	1:NO
238	2:YES	1:NO
239	2:YES	1:NO
240	1:NO	1:NO
241	1:NO	1:NO
242	2:YES	2:YES
243	2:YES	1:NO
244	2:YES	1:NO
245	1:NO	1:NO
246	1:NO	1:NO
247	1:NO	1:NO
248	2:YES	2:YES
249	1:NO	1:NO
250	1:NO	1:NO

251	1:NO	1:NO
252	1:NO	1:NO
253	1:NO	1:NO
254	1:NO	1:NO
255	1:NO	1:NO
256	1:NO	1:NO
257	1:NO	1:NO
258	1:NO	1:NO
259	1:NO	1:NO
260	1:NO	1:NO
261	1:NO	1:NO
262	1:NO	1:NO
263	1:NO	1:NO
264	2:YES	2:YES
265	1:NO	1:NO
266	1:NO	1:NO
267	2:YES	1:NO
268	1:NO	1:NO
269	1:NO	1:NO
270	1:NO	1:NO
271	1:NO	1:NO
272	1:NO	1:NO
273	2:YES	2:YES
274	1:NO	1:NO
275	1:NO	1:NO
276	1:NO	1:NO
277	2:YES	2:YES
278	1:NO	1:NO
279	1:NO	1:NO
280	1:NO	1:NO
281	1:NO	1:NO
282	1:NO	1:NO
283	1:NO	1:NO
284	1:NO	1:NO
285	1:NO	1:NO
286	1:NO	1:NO
287	1:NO	1:NO
288	1:NO	1:NO
289	1:NO	1:NO
290	1:NO	1:NO
291	1:NO	1:NO
292	1:NO	1:NO
293	1:NO	1:NO
294	1:NO	1:NO
295	2:YES	2:YES
296	1:NO	1:NO
297	1:NO	1:NO
298	1:NO	1:NO
299	1:NO	1:NO
300	1:NO	1:NO

301	2:YES	2:YES
302	1:NO	1:NO
303	1:NO	1:NO
304	1:NO	1:NO
305	1:NO	1:NO
306	1:NO	1:NO
307	1:NO	1:NO
308	1:NO	1:NO
309	2:YES	1:NO
310	1:NO	1:NO
311	1:NO	1:NO
312	1:NO	1:NO
313	1:NO	1:NO
314	1:NO	1:NO
315	1:NO	1:NO
316	1:NO	1:NO
317	1:NO	1:NO
318	1:NO	1:NO
319	1:NO	1:NO
320	1:NO	1:NO
321	1:NO	1:NO
322	2:YES	2:YES
323	2:YES	1:NO
324	1:NO	1:NO
325	1:NO	1:NO
326	1:NO	1:NO
327	1:NO	1:NO
328	1:NO	1:NO
329	1:NO	1:NO
330	1:NO	1:NO
331	1:NO	1:NO
332	1:NO	1:NO
333	1:NO	1:NO
334	2:YES	2:YES
335	1:NO	1:NO
336	1:NO	1:NO
337	1:NO	1:NO
338	1:NO	1:NO
339	1:NO	1:NO
340	1:NO	2:YES
341	1:NO	1:NO
342	1:NO	1:NO
343	1:NO	1:NO
344	1:NO	1:NO
345	1:NO	1:NO
346	1:NO	1:NO
347	1:NO	1:NO
348	1:NO	1:NO
349	2:YES	2:YES
350	1:NO	1:NO

351	1:NO	1:NO
352	1:NO	1:NO
353	1:NO	1:NO
354	1:NO	1:NO
355	1:NO	1:NO
356	1:NO	1:NO
357	1:NO	1:NO
358	1:NO	1:NO
359	1:NO	1:NO
360	1:NO	1:NO
361	1:NO	1:NO
362	1:NO	1:NO
363	1:NO	1:NO
364	1:NO	1:NO
365	1:NO	1:NO
366	1:NO	1:NO
367	1:NO	1:NO
368	1:NO	1:NO
369	1:NO	1:NO
370	1:NO	1:NO
371	1:NO	1:NO
372	1:NO	1:NO
373	1:NO	1:NO
374	1:NO	1:NO
375	1:NO	1:NO
376	1:NO	1:NO
377	1:NO	1:NO
378	2:YES	1:NO
379	1:NO	1:NO
380	1:NO	1:NO
381	1:NO	1:NO
382	1:NO	1:NO
383	1:NO	1:NO
384	1:NO	1:NO
385	1:NO	1:NO
386	1:NO	1:NO
387	2:YES	1:NO
388	1:NO	1:NO
389	1:NO	1:NO
390	1:NO	1:NO
391	1:NO	1:NO
392	2:YES	2:YES
393	1:NO	1:NO
394	1:NO	1:NO
395	1:NO	1:NO
396	2:YES	1:NO
397	1:NO	1:NO
398	1:NO	1:NO
399	1:NO	1:NO
400	2:YES	2:YES

401	1:NO	1:NO
402	1:NO	1:NO
403	2:YES	2:YES
404	1:NO	1:NO
405	1:NO	1:NO
406	2:YES	1:NO
407	2:YES	2:YES
408	1:NO	1:NO
409	1:NO	1:NO
410	1:NO	1:NO
411	1:NO	1:NO
412	2:YES	1:NO
413	1:NO	1:NO
414	2:YES	2:YES
415	1:NO	1:NO
416	1:NO	1:NO
417	1:NO	1:NO
418	1:NO	1:NO
419	1:NO	1:NO
420	1:NO	1:NO
421	1:NO	1:NO
422	1:NO	1:NO
423	1:NO	1:NO
424	1:NO	1:NO
425	1:NO	1:NO
426	1:NO	1:NO
427	1:NO	1:NO
428	1:NO	1:NO
429	1:NO	1:NO
430	1:NO	1:NO
431	1:NO	1:NO
432	2:YES	1:NO
433	1:NO	1:NO
434	1:NO	1:NO
435	1:NO	1:NO
436	1:NO	1:NO
437	1:NO	1:NO
438	1:NO	1:NO
439	1:NO	1:NO
440	2:YES	2:YES
441	1:NO	1:NO
442	1:NO	1:NO
443	1:NO	1:NO
444	1:NO	1:NO
445	2:YES	1:NO
446	2:YES	2:YES
447	1:NO	1:NO
448	1:NO	1:NO
449	1:NO	1:NO
450	1:NO	1:NO

451	1:NO	1:NO
452	1:NO	1:NO
453	1:NO	1:NO
454	1:NO	1:NO
455	1:NO	1:NO
456	2:YES	2:YES
457	1:NO	1:NO
458	1:NO	1:NO
459	1:NO	1:NO
460	1:NO	1:NO
461	1:NO	1:NO
462	1:NO	1:NO
463	1:NO	1:NO
464	1:NO	1:NO
465	1:NO	1:NO
466	1:NO	1:NO
467	1:NO	1:NO
468	1:NO	1:NO
469	1:NO	1:NO
470	1:NO	1:NO
471	2:YES	1:NO
472	1:NO	1:NO
473	1:NO	1:NO
474	1:NO	1:NO
475	1:NO	1:NO
476	1:NO	1:NO
477	1:NO	1:NO
478	1:NO	1:NO
479	1:NO	1:NO
480	1:NO	1:NO
481	1:NO	1:NO